









395  
79.70















# SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS

## CONTENTS OF VOLUME TWENTYONE, 1959

( All rights reserved )

### PAPERS

Relation between stature and blood group among Indian Soldiers. <i>By N. T. Mathew</i> .. .. .	1
A partial order and its applications to probability theory. <i>By T. V. Narayana</i>	91
Random processes in economic theory and analysis. <i>By P. A. P. Moran</i> ..	99
Expressions for the lower bound to confidence coefficients. <i>By Saibal Kumar Banerjee</i> .. .. .	127
A Pilot Health Survey in West Bengal—1955. <i>By S. J. Poti, M. V. Raman, S. Biswas and B. Chakraborty</i> .. .. .	141
On recall lapse in infant death reporting. <i>By Ranjan Kumar Som and Nitai Chandra Das</i> .. .. .	205
The analysis of heterogeneity. I. <i>By J. B. S. Haldane</i> .. .. .	209
Sufficient statistics of minimal dimension. <i>By Edward W. Barankin and Melvin Katz, Jr.</i> .. .. .	217
The family of ancillary statistics. <i>By D. Basu</i> .. .. .	247
Metricizing rank-ordered or unordered data for a linear factor analysis. <i>By Louis Guttman</i> .. .. .	257
Positive and negative dependence of two random variables. <i>By H. S. Konijn</i>	269
Definition and use of generalized percentage points. <i>By John E. Walsh</i> ..	281
Joint asymptotic distribution of $U$ -statistics and order statistics. <i>By J. Sethuraman and B. V. Sukhatme</i> .. .. .	289
Some sampling systems providing unbiased ratio estimators. <i>By N. S. Nanjamma, M. N. Murthy and V. K. Sethi</i> .. .. .	299
Tables for some small sample tests of significance for Poisson distributions and $2 \times 3$ contingency tables. <i>By I. M. Chakravarti and C. Radhakrishna Rao</i> .. .. .	315
Expected values of mean squares in the analysis of incomplete block experiments and some comments based on them. <i>By C. Radhakrishna Rao</i>	327
Some remarks on the missing plot analysis. <i>By Sujit Kumar Mitra</i> ..	337
The use of linear algebra in deriving prime power factorial designs with confounding and fractional replication. <i>By Norman T. J. Bailey</i> ..	345
Properties of the invariant $I_m$ ( $m$ -odd) for distributions admitting sufficient statistics. <i>By B. Raja Rao</i> .. .. .	355
Numerical evaluation of certain multivariate normal integrals. <i>By Peter Ihm</i>	363
On the evaluation of the probability integral of a multivariate normal distribution. <i>By S. John</i> .. .. .	367
The distribution of Wald's classification statistic when the dispersion matrix is known. <i>By S. John</i> .. .. .	371





An extension of Hald's table for the one-sided censored normal distribution. By <i>Nikhilesh Bhattacharya</i> .. .. .	377
Almost unbiased ratio estimates based on interpenetrating sub-sample estimates. By <i>M. N. Murthy and N. S. Nanjamma</i> .. .. .	381
Precision in the construction of cost of living index numbers. By <i>K. S. Banerjee</i>	393
Price indexes and sampling. By <i>Erland v. Hofsten</i> .. .. .	401

#### REPORTS

National Sample Survey : Number Eleven : The Sample Survey of Manu- facturing Industries, 1949 and 1950 .. .. .	13
National Sample Survey : Number Twelve : A Technical Note on Age Grouping	57

#### CORRIGENDA

Bias in estimation of serial correlation coefficients. By <i>A. Sree Rama Sastry</i>	404
Expressions for the lower bound to confidence coefficients. By <i>Saibal Kumar Banerjee</i> .. .. .	404

#### AUTHOR INDEX

<i>Bailey, Norman T. J.</i> The use of linear algebra in deriving prime power- factorial designs with confounding and fractional replication ..	345
<i>Banerjee, K. S.</i> Precision in the construction of cost of living index numbers ..	393
<i>Banerjee, Saibal Kumar.</i> Expressions for the lower bound to confidence coefficients .. .. .	127
<i>Barankin, Edward W. and Katz, Melvin, Jr.</i> Sufficient statistics of minimal dimension .. .. .	217
<i>Basu, D.</i> The family of ancillary statistics .. .. .	247
<i>Biswas, S., Poti, S. J., Raman, M. V. and Chakraborty, B.</i> A pilot health survey in West Bengal—1955 .. .. .	141
<i>Bhattacharya, Nikhilesh.</i> An extension of Hald's table for the one-sided censored normal distribution .. .. .	377
<i>Chakravarti, I. M. and Rao, C. Radhakrishna.</i> Tables for some small sample tests of significance for Poisson distributions and $2 \times 3$ contingency tables .. .. .	315
<i>Chakraborty, B., Poti, S. J., Raman, M. V. and Biswas, S.</i> A pilot health survey in West Bengal—1955 .. .. .	141
<i>Das, Nitai Chandra and Som, Ranjan Kumar.</i> On recall lapse in infant death reporting .. .. .	205
<i>Guttman, Louis.</i> Metricizing rank-ordered or unordered data for a linear factor analysis .. .. .	257
<i>Haldane, J. B. S.</i> The analysis of heterogeneity. I. .. .. .	209
<i>Hofsten, Erland v.</i> Price indexes and sampling .. .. .	401
<i>Ihm, Peter.</i> Numerical evaluation of certain multivariate normal integrals	363



<i>John, S.</i> On the evaluation of the probability integral of a multivariate normal distribution .. .. .	367
——— The distribution of Wald's classification statistic when the dispersion matrix is known .. .. .	371
<i>Katz, Melvin, Jr. and Barankin Edward W.</i> Sufficient statistics of minimal dimension .. .. .	217
<i>Konijn, H. S.</i> Positive and negative dependence of two random variables	269
<i>Mathew, N. T.</i> Relation between stature and blood group among Indian soldiers .. .. .	1
<i>Moran, P. A. P.</i> Random processes in economic theory and analysis ..	99
<i>Mitra, Sujit Kumar.</i> Some remarks on the missing plot analysis ..	337
<i>Murthy, M. N., Nanjamma, N. S., and Sethi, V. K.</i> Some sampling systems providing unbiased ratio estimators .. .. .	299
<i>Murthy, M. N. and Nanjamma, N. S.</i> Almost unbiased ratio estimates based on interpenetrating sub-sample estimates .. .. .	381
<i>Nanjamma, N. S., Murthy, M. N. and Sethi, V. K.</i> Some sampling systems providing unbiased ratio estimators .. .. .	299
<i>Nanjamma, N. S. and Murthy, M. N.</i> Almost unbiased ratio estimates based on interpenetrating sub-sample estimates .. .. .	381
<i>Narayana, T. V.</i> A partial order and its applications to probability theory ..	91
<i>Poti, S. J., Raman, M. V., Biswas, S. and Chakraborty, B.</i> A pilot health survey in West Bengal—1955 .. .. .	141
<i>Rao, C. Radhakrishna and Chakravarti, I. M.</i> Tables for some small sample tests of significance for Poisson distributions and $2 \times 3$ contingency tables .. .. .	315
<i>Rao, C. Radhakrishna.</i> Expected values of mean squares in the analysis of incomplete block experiments and some comments based on them ..	327
<i>Rao, B. Raja.</i> Properties of the invariant $I_m$ ( $m$ -odd) for distributions admitting sufficient statistics .. .. .	355
<i>Raman, M. V., Poti, S. J., Biswas, S. and Chakraborty, B.</i> A pilot health survey in West Bengal—1955 .. .. .	141
<i>Sethi, V. K., Nanjamma, N. S. and Murthy, M. N.</i> Some sampling systems providing unbiased ratio estimators .. .. .	299
<i>Sethuraman, J. and Sukhatme, B. V.</i> Joint asymptotic distribution of $U$ -statistics and order statistics .. .. .	289
<i>Som, Ranjan Kumar and Das, Nitai Chandra.</i> On recall lapse in infant death reporting .. .. .	205
<i>Sukhatme, B. V. and Sethuraman, J.</i> Joint asymptotic distribution of $U$ -statistics and order statistics .. .. .	289
<i>Walsh, John E.</i> Definition and use of generalized percentage points ..	281







# SANKHYĀ

## THE INDIAN JOURNAL OF STATISTICS

*Edited by : P. C. MAHALANOBIS*

---

VOL. 21, PARTS 1 & 2

MARCH

1959

---

### RELATION BETWEEN STATURE AND BLOOD GROUP AMONG INDIAN SOLDIERS

*By N. T. MATHEW*  
*Army Headquarters, India*

**SUMMARY.** This paper analyses blood group data relating to 4543 soldiers of the Indian army surveyed in 1952-53. It is seen that there are significant differences in stature among the different blood groups. Group B is the tallest and group A the shortest. These differences persist even when the data are considered separately for different states and communities. Analysis of the gene frequencies in different states and communities reveal certain interesting group affinities.

#### 1. INTRODUCTION

An important source of interest in the study of blood group frequencies is the light that they throw on anthropological differences. Anthropometrists had been using in their work measurements of body dimensions for a long time even before they started analysing blood groups. But no attempt seems to have been made to correlate blood groups and body measurements. Some indications that blood groups are related to physical traits such as proneness to contract certain diseases are referred to by Mourant (1954). The data used in the present paper reveal significant variation in stature among persons of different blood groups. This does not appear to have been noticed before.

#### 2. THE SAMPLE

The present data relate to 4543 soldiers of the Indian Army who formed part of a somewhat larger sample of soldiers selected for a survey of body measurements carried out in 1952-53. The primary object of the survey was the collection of data for standardization of clothing sizes. A medical officer, Capt. D. N. Bhattacharya, who was in charge of the field work found time for blood group determinations while his measuring team was busy on the body measurements.

Soldiers of the Indian Army cannot be regarded as a random sample of the general population of the country. The volunteers, who come forward for recruitment belong in varying proportions to different economic, social and regional strata. The actual recruits are further selected to conform to certain standards of height, weight and other physical characteristics.

The 4543 soldiers considered here do not constitute a random sample of soldiers. They were chosen from units located at the time of survey in Delhi and some



other stations, so as to obtain 200 soldiers from each of a number of 'army classes' which had to be studied separately for clothing sizes.

However, these considerations may not affect conclusions about blood group frequencies as blood group did not influence the selection in any way.

The 'states' referred to in this paper are the pre-reorganization states which existed in India at the time of the survey plus Nepal which is outside India. The communities are either tribes (e.g. Adibasis, Ahirs) or linguistic-territorial groups (e.g. Bengalees, Biharis, Tamilians) or religious groups (Muslims, Sikhs, Christians). None from these three religious groups are included in any of the tribal or linguistic or territorial groups. Statements made by the subjects at the time of the survey form the basis of grouping. 'Sikhs (M & R)' stand for Mazhabi and Ramdasia Sikhs who are supposed to have belonged originally to low caste Hindus. 'Syrian Christians' are an indigenous group whose connection with Syria is not racial.

### 3. DIFFERENCES IN STATURE

The main object of this paper is to invite attention to differences among the average values of height in persons belonging to the four ABO blood groups. In Table 1 we give the analysis of variance of height between and within blood groups.

TABLE 1. ANALYSIS OF VARIANCE OF HEIGHT (cm<sup>2</sup>)

source of variation	d.f.	s.s.	m.s.	F
between blood groups	3	467	155.6	4.37
within blood groups	4539	161607	35.6	
total	4542	162074		

The ratio of variances which comes out as 4.37 exceeds the one per cent level of significance. The mean height for each blood group is given in Table 2.

TABLE 2. MEANS AND STANDARD ERRORS OF HEIGHT IN CMS

blood groups	O	A	B	AB	total
number of observations	1480	1242	1406	415	4543
mean	167.6	167.2	168.0	167.3	167.5
standard error of mean	0.16	0.17	0.16	0.29	0.09

It would appear that the 'B' group is taller than the other phenotypes. Second in order of height comes 'O', third is 'AB' and the shortest is 'A'.

The significance of the variance ratio of Table 1 may possibly be due to the total sample being a mixture of individuals from different parts of India with different proportions of the O, A, B, AB phenotypic frequencies and different average heights. The individuals are classified by state and communities for a closer study in the following section.

## RELATION BETWEEN STATURE AND BLOOD GROUP

Similar analysis for weight as well as blood pressure did not reveal any significant differences. The averages of age for the four groups were also found to be nearly equal, being 26.7, 26.6, 26.6 and 26.7 respectively for O, A, B and AB. Even for height the magnitude of the difference is so small that significance could not have been achieved in smaller series of observations. This probably explains why such differences were not noticed before.

Analysis of variance was carried out separately for the twenty states and thirtytwo communities considered in this paper. The total numbers in these states and communities vary from 4 to 970. Only in one case did the variance ratio prove significant.

### 4. STATES AND COMMUNITIES

The effect, if any, of blood group on height must be regarded as superimposed on the effect of environmental and racial differences. These two latter factors may to some extent be reflected in the differences between states (Table 3) and communities (Table 4).

TABLE 3. AVERAGE HEIGHT BY BLOOD GROUPS AND STATES

state	frequency					average height in cm				total
	O	A	B	AB	total	O	A	B	AB	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Assam	78	84	42	5	209	160.8	161.1	161.7	160.9	161.1
Bihar	61	84	82	42	269	167.1	166.8	166.3	166.2	166.6
Bombay	116	104	117	36	373	165.7	165.8	166.0	166.1	165.8
Coorg	5	1	4	2	12	167.0	166.5	167.9	170.5	167.8
Delhi	8	6	8	—	22	170.5	172.5	176.6	—	173.3
Himachal Pradesh	7	9	8	2	26	168.2	169.1	170.0	171.5	169.3
Hyderabad	5	2	3	1	11	166.1	167.0	165.7	160.5	165.6
Jammu and Kashmir	67	83	97	28	275	168.5	169.3	168.4	169.0	168.7
Madhya Bharat	3	—	5	1	9	167.2	—	169.3	166.0	168.2
Madhya Pradesh	20	17	19	7	63	167.6	166.9	165.6	165.0	166.5
Madras	229	109	171	34	543	167.5	166.8	167.0	167.5	167.2
Mysore	6	3	2	3	14	163.8	164.3	168.5	171.0	166.1
Orissa	5	1	3	1	10	166.8	158.5	165.3	160.5	164.9
PEPSU	62	41	51	21	175	171.7	169.6	171.4	169.5	170.8
Punjab	291	262	329	88	970	170.1	170.3	170.5	169.3	170.2
Rajasthan	110	54	80	14	258	170.5	170.5	171.1	169.5	170.6
Travancore-Cochin	86	56	50	7	199	167.3	165.5	166.9	165.4	166.6
Uttar Pradesh	200	217	240	100	757	167.2	166.3	167.1	166.1	166.8
West Bengal	50	48	49	11	158	167.3	167.1	167.4	166.0	167.2
Nepal	71	61	46	12	190	161.0	162.3	162.2	163.0	161.9
total	1480	1242	1406	415	4543	167.6	167.2	168.0	167.3	167.5
number of group averages	}					9	7	16	7	
greater than the state averages						10	9.5	10	9.5	



It will be seen from Table 3 that among states the average height varies by 12.2 cms from 161.1 cms in Assam to 173.3 cms in Delhi. Yet in 16 out of 20 states the average height of B is greater than the general average for the state. Due to chance only 10 out of 20 states can be expected to have B taller than the average. The difference between the numbers observed and expected can be seen to be statistically significant.

TABLE 4. AVERAGE HEIGHT BY BLOOD GROUPS AND COMMUNITIES

community	frequency				total	average height in cms				total
	O	A	B	AB		O	A	B	AB	
Adibasis (Bihar)	27	37	41	21	126	163.4	164.7	164.5	163.4	164.1
Adibasis (Other)	8	12	10	2	32	164.7	163.8	165.8	160.5	164.4
Ahirs	77	62	74	24	237	171.5	171.4	172.5	171.2	171.8
Andhras	36	21	29	7	93	167.5	164.8	165.7	167.0	166.3
Assamese	76	38	81	5	200	160.5	161.0	160.8	160.9	160.8
Balmikis	4	5	17	4	30	163.4	168.8	166.7	159.0	165.6
Bengalees	53	53	57	12	175	168.0	166.9	167.5	165.3	167.3
Biharis	22	29	27	18	96	171.1	170.0	169.1	170.4	170.3
Christians (Syrian)	37	22	23	2	84	167.2	165.4	167.3	169.3	166.8
Christians (Tamil)	19	13	29	6	67	167.3	168.7	166.7	164.1	167.0
Christians (Other)	5	1	5	—	11	163.0	170.5	162.9	—	163.7
Coorgs	5	1	4	1	11	167.0	166.5	167.9	172.0	167.7
Dogras	59	97	74	38	268	168.1	168.2	168.8	168.4	168.4
Garhwalis	40	75	57	24	196	162.5	163.0	163.9	163.0	163.2
Gurkhas	73	64	49	14	200	160.9	162.4	162.4	162.9	161.9
Gujjars	62	48	79	11	200	170.8	170.0	171.0	171.0	170.7
Hindus (U.P.)	17	19	23	19	78	167.8	170.2	165.1	167.4	167.5
Jats	51	31	46	9	137	171.5	171.7	171.7	173.6	171.8
Jammu Hindus	50	55	77	18	200	168.4	168.2	168.1	169.5	168.3
Kanarese	3	2	1	2	8	171.7	166.0	171.5	170.8	170.0
Kumaonis	50	64	57	29	200	166.7	166.6	165.7	165.2	166.2
Lingayats	8	4	4	2	18	165.4	167.6	167.8	169.5	166.9
Mahars	64	51	64	21	200	164.4	163.9	164.6	165.8	164.5
Marathas	63	64	63	18	208	167.0	167.5	167.1	165.9	167.1
Malayalees	105	69	53	13	240	167.2	166.3	167.4	167.0	167.0
Muslims (U.P.)	15	9	20	4	48	166.0	166.2	169.7	165.0	167.5
Oriyas	1	—	2	1	4	163.0	—	164.3	160.5	163.0
Punjabis	38	33	39	14	124	168.9	170.7	170.5	168.5	169.8
Rajputs	126	67	105	22	320	170.0	170.9	170.0	168.8	170.1
Sikhs (M & R)	66	52	67	17	202	167.6	167.6	168.5	167.0	167.9
Sikhs (Other)	100	59	88	21	268	172.6	173.1	173.0	171.1	172.7
Tamilians	120	42	84	16	262	167.6	166.7	167.3	168.1	167.4
total	1480	1242	1406	415	4543	167.6	167.2	168.0	167.3	167.5
number of group averages greater than the community average	}				observed	13.5	12	22	14	
					expected	16	15.5	16	15.5	

Similarly it can be seen from Table 4 that the average height varies from 160.8 among Assamese to 172.7 among Sikhs (Other). But here also against an



## RELATION BETWEEN STATURE AND BLOOD GROUP

expected number of 16 communities we have actually 22 communities in which the 'B' group is taller than the average. The difference can be seen to be significant at the 5 per cent level.

From the evidence considered above, it seems likely that there are significant though small differences in stature associated with blood groups. Group 'B' has the highest average stature and probably 'A' the lowest. It is not easy to explain why this should be so. This is unlikely to be the result of intermixture with a tall race which came into India bringing with it also a high percentage of the B gene. It is found that some of the primitive tribes of India have high proportion of B. Though it is true that the races of Central Asia have comparatively higher frequency of B, we have no evidence that they were also tall. The 'Aryans' are believed to have come into India from the direction of Persia. But the Aryans probably had a low frequency of B as is the case with present day populations in some Western European countries.

The only tenable theory would be to regard a contribution to stature as the effect of the B gene itself or some closely linked gene. Height is known to be the result of a large number of genes. Perhaps one or two of these are linked to the B gene.

### 5. GENE FREQUENCIES

In Tables 5 and 6 gene frequencies estimated by Bernstein's method are given respectively for the states and for the communities. Only such states and communities are shown as are represented by more than 25 individuals in our sample. Charts 1 and 2 give graphical representations of the gene frequencies in Tables 5 and 6 by means of trilinear coordinates.

TABLE 5. DISTRIBUTION OF BLOOD GROUPS BY STATES

state	frequency of phenotype				total	gene percentage			$\chi^2$ 1 d.f.
	O	A	B	AB		p	q	r	
Assam	78	84	42	5	209	24.55	12.12	63.30	6.79**
Bihar	61	84	82	42	269	26.91	26.41	46.67	0.71
Bombay	166	104	117	36	373	20.27	23.21	55.82	0.00
Himachal Pradesh	7	9	8	2	26	24.35	21.82	53.81	0.37
Jammu and Kashmir	67	83	97	28	275	22.97	26.37	50.65	1.54
Madhya Pradesh	20	17	19	7	63	21.21	23.24	55.54	0.18
Madras	229	109	171	34	543	14.16	21.08	64.76	0.12
Nepal	71	61	46	12	190	21.60	16.71	61.69	0.33
PEFSU	62	41	51	21	175	19.40	23.00	57.58	3.03
Punjab	291	262	329	88	970	20.12	24.58	55.30	1.10
Rajasthan	110	54	80	14	258	14.20	20.29	65.50	0.08
Travancore-Cochin	86	56	50	7	199	17.45	15.63	66.91	1.99
Uttar Pradesh	200	217	240	100	757	23.65	25.66	50.66	1.27
West Bengal	50	47	49	12	158	20.97	21.78	57.24	0.67
other states	32	14	25	7	78	14.38	23.00	62.60	1.03
total	1480	1242	1406	415	4543	20.30	22.60	57.11	0.00

\*\*significant at 1% level



TABLE 6. DISTRIBUTION OF BLOOD GROUPS BY COMMUNITIES

community	frequency of phenotype				total	gene percentage			$\chi^2$ 1 d.f.
	O	A	B	AB		p	q	r	
Adibasis (Bihar)	27	37	41	21	126	26.33	28.51	45.16	0.45
Adibasis (Other)	8	12	10	2	32	25.51	21.37	53.08	1.12
Ahirs	77	62	74	24	237	20.12	23.35	56.53	0.22
Andhras	36	21	29	7	93	16.37	21.68	61.95	0.03
Assamese	76	81	38	5	200	24.80	11.54	63.64	5.52*
Balmikis	4	5	17	4	30	16.49	45.66	37.84	0.14
Bengalees	53	52	57	13	175	20.89	22.73	56.37	1.32
Biharis	22	29	27	18	96	28.05	26.64	45.28	1.82
Christians (Syrian)	37	22	23	2	84	15.64	16.35	68.01	1.76
Christians (Tamil)	19	13	29	6	67	15.40	30.97	53.64	0.04
Dogras	59	97	74	38	268	29.53	23.68	46.79	0.02
Garhwalis	40	75	57	24	196	29.91	23.61	46.47	0.93
Gujjars	62	48	79	11	200	16.23	26.15	57.60	3.48
Gurkhas	73	64	49	14	200	21.95	17.27	60.78	0.15
Hindus (U.P.)	18	19	23	19	79	26.94	30.42	42.51	5.80*
Jammu Hindus	49	55	77	18	199	20.66	28.02	51.31	1.99
Jats	52	31	46	9	138	15.77	22.50	61.73	0.10
Kumaonis	49	64	57	29	199	26.84	24.49	48.67	0.57
Marathas	63	64	63	18	208	22.27	21.96	55.76	0.46
Mahars	64	51	64	21	200	19.93	24.08	55.99	0.30
Malayalees	105	69	53	13	240	18.88	14.86	66.26	0.02
Muslims (U.P.)	15	9	20	4	48	14.62	29.32	56.06	0.00
Punjabis	38	33	39	14	124	21.10	24.22	54.67	0.24
Rajputs	126	67	105	22	320	15.02	22.33	62.65	0.01
Sikhs (M & R)	66	52	67	17	202	18.90	23.62	57.48	0.11
Sikhs (other)	100	59	88	21	268	16.22	22.94	60.84	0.10
Tamilians	120	42	84	16	262	11.72	21.28	67.00	0.96
other communities	22	9	16	5	52	14.34	22.52	63.12	1.22
total	1480	1242	1406	415	4543	20.30	22.60	57.11	0.00

\*significant at 5% level

The last column in Tables 5 and 6 gives value of  $\chi^2$  (one degree of freedom) for testing the agreement between the observed numbers of phenotypes, and the expected numbers calculated from the estimated values of  $p$ ,  $q$  and  $r$ . The agreement is satisfactory except for Assamese and for U.P. Hindus.

# RELATION BETWEEN STATURE AND BLOOD GROUP

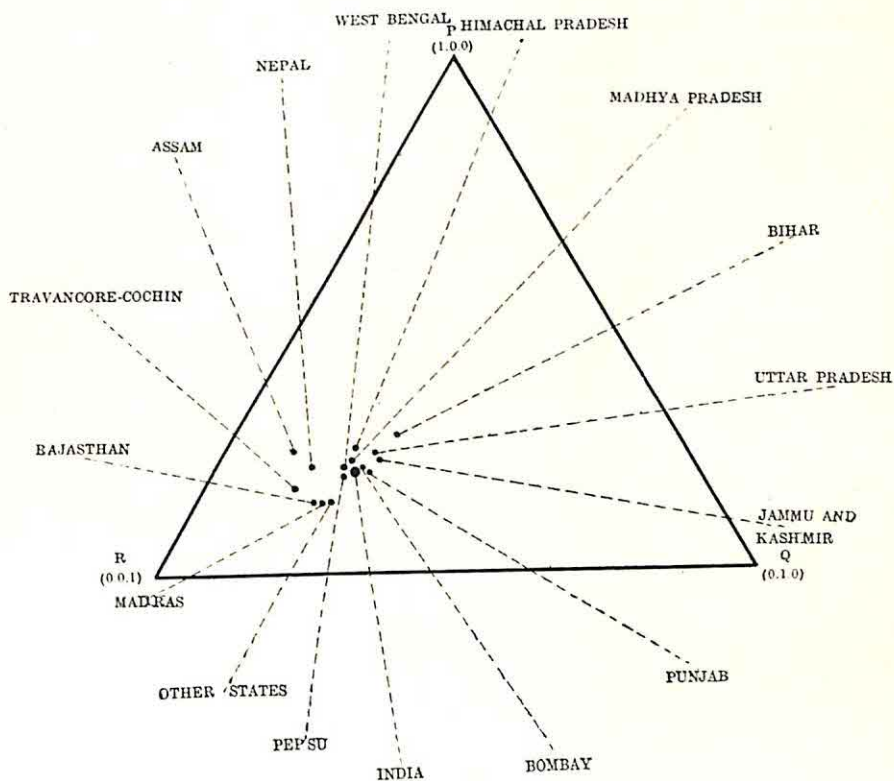


Chart 1. Blood group gene frequencies in different states.

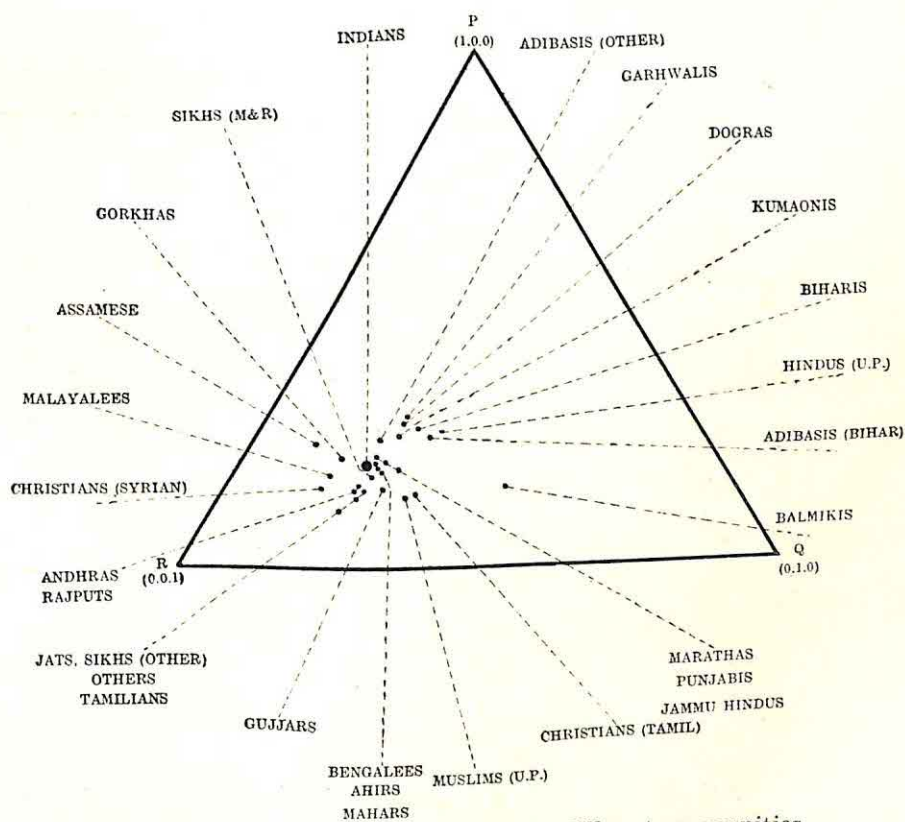


Chart 2. Blood group gene frequencies in different communities.



It is seen from Chart 1 that the populations of the States of Jammu and Kashmir, Punjab, PEPSU, Himachal Pradesh, Uttar Pradesh, West Bengal, Bombay and perhaps Bihar form a homogeneous group. Geographically, this group of states forms a fairly compact region stretching across North and Central India. Assam, Nepal and Travancore-Cochin are distinct from this group and from each other but all are on the side of low 'B' gene frequency. Rajasthan and Madras are surprisingly close to each other.

Chart 2 shows that some of the primitive tribes of India are comparatively rich in 'B' genes. Balmikis constitute a notable illustration with the highest percentage of 'B' genes and lowest O. The lowest frequency of B is among Assamese, Gorkhas and Malayalees. The Malayalee Hindus and Malayalee Syrian Christians appear to be racially close to each other, whereas the distance between the Tamil Hindus and Tamil Christians is considerable. The U.P. Muslims and U.P. Hindus also seem to be distinct though the percentage of the 'B' genes in both groups is nearly same.

Detailed figures of blood group is given in Table 7. We have not calculated gene frequencies from the detailed figures in Table 7 as the total numbers are small in most cases.

TABLE 7. DISTRIBUTION OF SOLDIERS ACCORDING TO BLOOD GROUP BY COMMUNITIES WITHIN EACH STATE

state	community	frequencies				total
		O	A	B	AB	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Assam	Assamese	76	81	38	5	200
	Bengalees	2	3	4	—	9
	total	78	84	42	5	209
Bihar	Adibasis (Bihar)	27	37	41	21	126
	Adibasis (Other)	5	11	9	2	27
	Ahirs	3	3	2	—	8
	Bengalees	1	1	1	1	4
	Biharis	22	29	27	18	96
	Gurkhas	—	—	1	—	1
	Rajputs	2	3	1	—	6
	Sikhs (Other)	1	—	—	—	1
	total	61	84	82	42	269
Bombay	Balmikis	—	—	1	—	1
	Kanarese	—	1	—	—	1
	Lingayats	6	4	4	2	16
	Mahars	54	41	57	19	171
	Marathas	55	56	55	15	181
	Punjabis	—	1	—	—	1
	Rajputs	—	1	—	—	1
	Sikhs (Other)	1	—	—	—	1
	total	116	104	117	36	373

# RELATION BETWEEN STATURE AND BLOOD GROUP

TABLE 7. DISTRIBUTION OF SOLDIERS ACCORDING TO BLOOD GROUP BY COMMUNITIES WITHIN EACH STATE (Continued)

state	community	frequencies				total
		O	A	B	AB	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Coorg	Coorgs	5	1	4	1	11
	Kanarese	—	—	—	1	1
	total	5	1	4	2	12
Delhi	Ahirs	1	2	2	—	5
	Balmikis	1	—	—	—	1
	Gujjars	—	2	—	—	2
	Jats	1	—	4	—	5
	Punjabis	4	1	1	—	6
	Rajputs	—	1	—	—	1
	Sikhs (Other)	1	—	1	—	2
	total	8	6	8	—	22
Himachal Pradesh	Dogras	5	9	8	2	24
	Punjabis	1	—	—	—	1
	Rajputs	1	—	—	—	1
	total	7	9	8	2	26
Hyderabad	Andhras	2	—	—	1	3
	Christians (Tamil)	—	1	2	—	3
	Lingayats	1	—	—	—	1
	Mahars	1	1	1	—	3
	Tamilians	1	—	—	—	1
	total	5	2	3	1	11
Jammu & Kashmir	Dogras	15	24	14	9	62
	Jats	—	1	1	—	2
	Jammu Hindus	50	55	77	18	200
	Punjabis	—	1	—	—	1
	Rajputs	—	1	2	—	3
	Sikhs (other)	2	1	3	1	7
	total	67	83	97	28	275
Madhya Bharat	Ahirs	—	—	—	1	1
	Gujjars	1	—	1	—	2
	Muslims (U.P.)	—	—	1	—	1
	Rajputs	2	—	3	—	5
	total	3	—	5	1	9
Madhya Pradesh	Adibasis	1	—	—	—	1
	Bengalees	—	—	3	1	4
	Christians (Tamil)	1	—	—	—	1
	Gujjars	1	—	—	—	1
	Mahars	8	9	6	2	25
	Marathas	8	8	7	3	26
	Punjabis	1	—	1	—	2
	Rajputs	—	—	2	1	3
	total	20	17	19	7	63



TABLE 7. DISTRIBUTION OF SOLDIERS ACCORDING TO BLOOD GROUP BY COMMUNITIES WITHIN EACH STATE (*Continued*)

state	community	frequencies				total
		O	A	B	AB	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Madras	Andhras	31	20	28	5	84
	Christians (Other)	5	1	5	—	11
	Christians (Syrian)	6	2	4	1	13
	Christians (Tamil)	18	11	26	6	61
	Kanarese	3	—	1	—	4
	Malayalees	52	36	24	7	119
	Marathas	—	—	1	—	1
	Tamilians	114	39	82	15	250
	total	229	109	171	34	543
Mysore	Andhras	1	1	—	1	3
	Kanarese	—	1	—	1	2
	Lingayats	1	—	—	—	1
	Mahars	1	—	—	—	1
	Tamilians	3	1	2	1	7
	total	6	3	2	3	14
Orissa	Adibasis	2	1	1	—	4
	Andhras	2	—	—	—	2
	Oriyas	1	—	2	1	4
	total	5	1	3	1	10
PEPSU	Ahirs	24	12	20	7	63
	Dogras	—	1	—	—	1
	Gujjars	2	1	3	—	6
	Jats	9	4	3	2	18
	Punjabis	1	1	1	2	5
	Rajputs	2	6	10	1	19
	Sikhs (M & R)	8	12	8	6	34
	Sikhs (Other)	16	4	6	3	29
	total	62	41	51	21	175
Punjab	Ahirs	25	28	34	8	95
	Balmikis	1	3	6	3	13
	Dogras	39	63	52	27	181
	Gurkhas	—	1	—	—	1
	Gujjars	18	16	28	4	66
	Jats	30	16	24	5	75
	Punjabis	30	27	34	11	102
	Rajputs	11	14	16	2	43
	Sikhs (M & R)	58	40	59	11	168
	Sikhs (Other)	79	54	76	17	226
	total	291	262	329	88	970

# RELATION BETWEEN STATURE AND BLOOD GROUP

TABLE 7. DISTRIBUTION OF SOLDIERS ACCORDING TO BLOOD GROUP BY COMMUNITIES WITHIN EACH STATE (Continued)

state	community	frequencies				total
		O	A	B	AB	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rajasthan	Ahirs	8	5	7	2	22
	Balmikis	—	—	1	—	1
	Bengalees	1	—	—	—	1
	Gujjars	23	19	28	3	73
	Hindus (U.P.)	—	1	—	—	1
	Jats	11	8	12	1	32
	Rajputs	67	21	32	8	128
	total	110	54	80	14	258
Travancore & Cochin	Bengalees	—	—	1	—	1
	Christians (Syrian)	31	20	19	1	71
	Christians (Tamil)	—	1	1	—	2
	Malayalees	53	33	29	6	121
	Tamilians	2	2	—	—	4
	total	86	56	50	7	199
U.P.	Ahirs	16	12	9	6	43
	Balmikis	2	2	9	1	14
	Bengalees	—	2	—	—	2
	Garhwalis	40	75	57	24	196
	Gurkhas	1	1	2	1	5
	Gujjars	17	10	19	4	50
	Hindus (U.P.)	17	18	23	19	77
	Jats	—	2	2	1	5
	Kumaonis	50	64	57	29	200
	Muslims (U.P.)	15	9	19	4	47
	Punjabis	1	2	2	1	6
	Rajputs	41	20	39	10	110
	Sikhs (Other)	—	—	2	—	2
	total	200	217	240	100	757
West Bengal	Andhras	—	—	1	—	1
	Bengalees	49	47	48	10	154
	Gurkhas	1	1	—	1	3
	total	50	48	49	11	158
Nepal	Gurkhas	71	61	46	12	190
grand total		1480	1242	1406	415	4543

## 6. COMPARISON WITH PREVIOUSLY PUBLISHED DATA

Mourant (1954) has quoted figures of blood group frequencies supplied by House and Mahalanobis (1953) for a number of groups based on data collected from the Indian Army during the 1939-45 War. Some of these groups are comparable with corresponding groups in the present paper. Relevant figures are given in Table 8. Majumdar and Bahadur (1952) have listed a large number of Indian groups for which blood group data have been published. The groups, Jats and Rajputs, quoted



by these authors appear to be comparable with corresponding groups in the present paper. These figures are also shown in Table 8.

TABLE 8. COMPARISON OF GENE PERCENTAGES

group	source of data	number tested	gene percentages		
			<i>p</i>	<i>q</i>	<i>r</i>
(1)	(2)	(3)	(4)	(5)	(6)
Punjab Hindus	present paper	124	21.10	24.22	54.67
	House and Mahalanobis	615	18.05	25.94	56.01
Rajputs	present paper	320	15.02	22.33	62.65
	House and Mahalanobis	111	17.18	25.76	57.06
	Malone and Lahiri	118	19.60	25.22	55.18
U. P. Hindus	present paper	79	26.94	30.42	42.51
	House and Mahalanobis	838	19.12	23.94	56.94
U. P. Muslims	present paper	48	14.62	29.32	56.06
	House and Mahalanobis	109	17.98	26.16	55.87
Jats	present paper	138	15.77	22.50	61.73
	Malone and Lahiri	277	17.28	24.14	58.58

The agreement between comparable gene percentages seem to be tolerably good except perhaps in the U.P. Hindus. The number of U.P. Hindus in the present sample is small and, moreover, there are a large number of castes in U.P. all of which may not have been represented in our sample. Majumdar and Rao (1958) give blood group data for a number of Bengal groups. The gene frequencies given by these authors in Table 9, p. 321 of their paper are consistent with the frequencies given in our Tables 6 and 7 for West Bengal and Bengalees.

I must thank Dr. C. R. Rao for many helpful suggestions in the preparation of this paper.

#### REFERENCES

- HOUSE, R. J. AND MAHALANOBIS, P. C. (1953): Personal communication referred to by Mourant (1954).  
 MAJUMDAR, D. N. AND BAHADUR, S. (1952): ABO Blood in India. *The Eastern Anthropologist*, 2 and 3, 101-122.  
 MAJUMDAR, D. N. AND RAO, C. R. (1958): Bengal anthropometric survey, 1945: A statistical study. *Sankhyā*, 19, 201-408.  
 MALONE AND LAHIRI, referred to by Majumdar and Bahadur (1952).  
 MOURANT, A. E. (1954): *The Distribution of the Human Blood Groups*, Blackwell Scientific Publications, Oxford.

*Paper received : November, 1958.*

# THE NATIONAL SAMPLE SURVEY

## NUMBER 11

### REPORT ON

### THE SAMPLE SURVEY OF MANUFACTURING INDUSTRIES

### 1949 AND 1950

#### FOREWORD

0.1 The National Income Committee\*, which was set up by the Government of India in 1949, found that the coverage of the Indian Census of Manufacturing Industries (which had been initiated annually from 1946 under the Industrial Statistics Act of 1942) was incomplete in many respects. It did not cover Part B and Part C States ; and excluded 34 out of 63 groups of industries into which all factories were divided for the purposes of the Census. Further, although by that time a larger number of establishments should have been considered as factories according to the 1948 Factories Act, the Census had been working on the basis of the older definition of factories according to the 1934 Act. The difference was indeed large. Under the 1948 Act there were about 28,000 factories in the country in 1949 but according to the older definition there were only about 17,000 factories ; and the Census was covering between 6,500 and 7,000 factories only.

0.2 The National Income Committee felt that it was very important for its work to have fairly reliable estimates of the contribution of the manufacturing industries to the national income as early as possible. On the recommendation of the Committee, the Government of India agreed to a quick survey on a sample basis being carried out by the Directorate of Industrial Statistics with the technical collaboration of the Indian Statistical Institute. For immediate needs, it was decided to have a sample survey of factories, as defined under the 1934 Act, in the first instance. The size of the sample was 1742 ; and information on a brief schedule was collected directly by investigators who visited the sample factories. The survey started in January 1951 and was completed in June next.

0.3 The survey was arranged in two instalments, and preliminary estimates based on the first instalment of the data, processed by the Indian Statistical Institute, were furnished to the Committee by April 1951, within four months from the commencement of the survey. The final estimates on the full material were made available within six months after the completion of the survey.

---

\* consisting of Professor P. C. Mahalanobis (Chairman), Professor D. R. Gadgil and Dr. V. K. R. V. Rao (Members), and Dr. R. C. Desai and later Sri Mani Mohan Mukherjee (Secretary).



0.4 The field schedule, which has been reproduced at the end of the report, was simple and was designed to supply information on the number of persons engaged, wages and salaries paid, and the net value added which are of basic importance for studying the trend of industrial activities and the growth of national income.

0.5 This survey had demonstrated the feasibility of using the sampling method on a 'voluntary' basis by the method of interview by investigators and its capacity to supply useful results quickly and at a low cost. Since then a Sample Survey of Manufacturing Industries with coverage of factories, in accordance with the 1948 Act, is being carried out every year, and it is intended to publish the results regularly in future.

29 August 1958

P. C. MAHALANOBIS

# THE NATIONAL SAMPLE SURVEY

## NUMBER 11

### REPORT ON THE SAMPLE SURVEY OF MANUFACTURING INDUSTRIES 1949 AND 1950

#### CONTENTS

	PAGE
FOREWORD : by P. C. Mahalanobis ... ..	13
CHAPTER ONE : Introduction ... ..	17
CHAPTER TWO : Coverage of the survey ... ..	19
CHAPTER THREE : Sampling design and organisation of work ...	21
CHAPTER FOUR : Reliability of estimates ... ..	24
CHAPTER FIVE : Summary results ... ..	27
APPENDIX I : Industry-wise table showing the number of sample factories and the total number of factories covered in SSMI : 1949 and 1950	48
APPENDIX II : Facsimile of the schedule of investigation ... ..	50
APPENDIX III : Principal participants ... ..	52

#### INDEX TO TABLES IN THE TEXT

##### CHAPTER THREE

TABLE 3.1 : List of industries with the number of establishments in each of them ..	21
---	----

##### CHAPTER FOUR

TABLE 4.1 : Comparison between sample survey and census results : 1949 and 1950 ..	24
TABLE 4.2 : Estimates of selected items as obtained from the two parts of the sample ..	25



## CHAPTER FIVE

	PAGE
TABLE 5.1 : Estimates of value of some selected items relating to manufacturing industries of India in 1949 and 1950 .. .. .	7
TABLE 5.2 : A few selected items relating to manufacturing industries in 1949 and 1950 ..	28
TABLE 5.3 : Estimates of selected items for some industry groups in 1949 and 1950 ..	29
TABLE 5.4 : Estimates of fixed capital and output per worker for some industry groups in 1949 and 1950 .. .. .	30
TABLE 5.5 : Total number of factories and the number of sample factories (relating to ten major manufacturing industries) in 1949 and 1950 .. .. .	31
TABLE 5.6 : Estimates of fixed and working capital and rent paid on fixed assets (in ten major manufacturing industries) in 1949 and 1950 .. .. .	32
TABLE 5.7 : Production account of 61 manufacturing industries in 1949 and 1950 ..	33
TABLE 5.8 : Estimates of cost items .. .. .	34
TABLE 5.9 : Increase or decrease in costs of materials etc. in 1950 over 1949 ..	35
TABLE 5.10 : Estimates of output for all industries in 1949 and 1950 ..	36
TABLE 5.11 : Percentage increase or decrease in the value of output in 1950 over 1949—ten major industries .. .. .	36
TABLE 5.12 : Value added by manufacture in 1949 and 1950—all industries ..	37
TABLE 5.13 : Value added by manufacture in 1949 and 1950—ten major industries ..	38
TABLE 5.14 : Estimates of fixed and working capital for all industries in 1949 and 1950 ..	39
TABLE 5.15 : Working capital as percentage of invested capital in 1949 and 1950—ten major industries .. .. .	39
TABLE 5.16 : Gross and net ratios of output to invested capital in 1949 and 1950 ..	40
TABLE 5.17 : Total and working population compared to the working population in industries ..	41
TABLE 5.18 : Estimates of wages and salaries paid .. .. .	43
TABLE 5.19 : Estimates of wages and salaries for workers and persons other than workers in 1949 and 1950 .. .. .	43
TABLE 5.20 : Number of workers in relation to total wages and value of input and output in 1949 and 1950 .. .. .	44
TABLE 5.21 : Number of persons employed, per capita cost of employment and gross output per employed person in all industries in 1949 and 1950 .. .. .	45
TABLE 5.22 : Capital and output per employed person in 1949 and 1950 .. .. .	46

# THE NATIONAL SAMPLE SURVEY

## NUMBER 11

### REPORT ON

### THE SAMPLE SURVEY OF MANUFACTURING INDUSTRIES

### 1949 AND 1950

*This report on the Sample Survey of Manufacturing Industries, 1949 and 1950 was prepared by the Indian Statistical Institute and is being published in the form in which it was submitted to the Government of India. The views contained in the report are not necessarily those of the Government of India.\**

#### CHAPTER ONE

#### INTRODUCTION

1.1. This report presents results of the Survey of Indian Manufactures undertaken for the first time on a sampling basis and covers the calendar years 1949 and 1950.

1.2. The Survey was conducted to collect certain statistics for the use of the National Income Committee and make them available within a very short period.

1.3. For the purpose of the Census of Manufacturing Industries (CMI), the industries have been divided into 63 groups. The census of manufactures covers only 29 out of these 63 groups. This report, however, gives estimates in respect of all the groups of Indian industries, except Railway workshops, repair shops and locomotive shops (CMI-58), and arms, ammunition and explosives (CMI-59) which were excluded from the present survey. The estimates in the report relate to the details of number of sample factories covered, fixed and working capital, employment, wages and salaries, materials consumed, products manufactured and value added by manufacture.

1.4. The work of planning the survey began in December 1950. As the National Income Committee wanted estimates by April 1951 for their preliminary report it was decided to divide the samples roughly into two equal parts. The field work in respect of the first part started by the middle of January 1951 and was

---

\*The draft report (Number D. 10) was submitted to the Government of India in November 1956.



completed by the third week of March. Preliminary estimates of the contribution of manufacturing industries to national income were furnished to the National Income Committee by April 1951.

1.5. After the survey of the first part was over, the field work in the second part was taken up by the same set of investigators. The establishments included in this part were covered by the middle of June 1951. Final estimates were made available to the National Income Committee by the end of the year and all the particulars based on both the parts taken together were analysed and the tables completed by February 1952.

CHAPTER TWO

COVERAGE OF THE SURVEY

2.1. The National Income Committee set up by the Ministry of Finance, Government of India, wanted statistics relating to manufacturing industries for estimating the contribution from large scale industries to national income. The figures for two calendar years, namely, 1949 and 1950 were wanted and a view was expressed that it would be convenient if some provisional figures could be made available by April 1951.

2.2. Although the Directorate of Industrial Statistics, Ministry of Commerce and Industry, Government of India, was conducting annual census of manufacturing industries, the lag between the completion of analysis and the years to which the data related was about 2 to 3 years. Therefore, when these figures for 1949 and 1950 were wanted by the National Income Committee, the years for which the CMI figures were available went up to only 1948. The chance of obtaining figures relating to 1949 or 1950 on the basis of complete census by April 1951 was very remote indeed. The figures given in the censuses of manufactures were wanting in another respect also. The censuses were covering only 29 out of 63 groups of industries located in part A States, some of the important part B States and a few part C States. For national income purposes, larger coverage, both in respect of industries and in respect of geographical area, was naturally considered desirable.

2.3. Accordingly, a special inquiry on a random sampling basis to cover all the 63 industries in all States was planned and arrangements were made to obtain the analysed results quickly. The Government of India, at the instance of the Chairman of the National Income Committee, sanctioned a scheme for this sample survey as an experiment. The Director of Industrial Statistics was made responsible for the organisation of the survey.

2.4. The questionnaire included the following groups of items and altogether there were 37 different items for each of the years in respect of each establishment :

- (i) value of fixed capital which included land and building, plant and machinery and other fixed assets;
- (ii) value of working capital which included stocks of fuel and raw materials, stocks of products and by-products and partly finished products and cash in hand and at banks;
- (iii) rent of fixed assets secured on lease;
- (iv) duration of working period;
- (v) labour employed with various breakdowns, and wages and salaries paid to them;



- (vi) value and quantity of input which included value of fuels, electricity, raw materials, chemicals and work done by other concerns; and
- (vii) value and quantity of output which included the value of products and by-products, and work done by the factory for customers.

2.5 The definitions used for the items were the same as those used for the Census of Manufactures.

2.6. As in the case of Census of Manufactures this survey was limited to manufacturing establishments employing 20 or more workers and using power. But the scope of the survey was extended to all States of the Indian Union and to all factories which come under Section 2(j) of the Factories Act, 1934 except two Government-run industries\*, *i.e.*, CMI-58 and CMI-59. The aggregate of all such manufacturing establishments was 17,377 exclusive of two industries mentioned above, according to the lists available with the Chief Inspectors of Factories of the different States.

---

\* Some data in respect of these two industries are available from the Railway Board and the Chief Statistical Officer, Army Headquarters respectively.

## CHAPTER THREE

## SAMPLING DESIGN AND ORGANISATION OF WORK

3.1. The frame for sampling consisted of a classified list of factories in India. Every manufacturing establishment in India employing 10 or more workers with power and 20 or more workers without power is required to be registered under the Indian Factories Act 1948. These establishments are registered with the Chief Inspectors of Factories of different States. The frame for the present survey was, however, restricted only to establishments employing 20 or more persons using power because no list was easily available which included the smaller establishments. While collecting the lists from the Chief Inspectors of different States the names and addresses of occupiers and the number of workers employed in establishments were also collected.

3.2. For convenience, a few of the 61 industries actually surveyed were further sub-divided and the total number, taking account of the sub-divisions, came to 69. Within each of these 69 industries, the establishments were classified into a number of groups according to the number of workers employed. For a number of industries which showed marked concentration in particular areas, establishments falling under any size-class were further grouped according to States. Thus, there were altogether 589 strata into which the establishments were classified.

3.3. Sample establishments were selected at random with equal probability from each of these strata and the total of samples was 1,885. The samples were allocated to the different strata in proportion to the total number of workers employed. Although the overall sampling fraction was approximately 1 in 9, the fraction between the different strata varied considerably. In eight industries, because the number of establishments was very small, all units were included in the sample. These eight industries are shown in Table (3.1).

TABLE (3.1) : LIST OF INDUSTRIES WITH THE NUMBER OF ESTABLISHMENTS IN EACH OF THEM

industry	c.m.i. number	number of establishments
(1)	(2)	(3)
1. sugar : gur and jaggery refineries	5(b)	7
2. aluminium, copper and brass : primary producers	22(a)	3
3. iron and steel : primary producers	23(a)	5
4. sewing machines	25	6
5. producer gas plants	26	3
6. electric lamps	27	10
7. turpentine and resin	37	2
8. petroleum refining *	39	1

\* return not received





3.4. The total sample was then divided roughly into two equal parts. The establishments under each part were scattered as widely as possible but the two parts were in effect not comparable sub-samples. After the survey of the first part was over, the second part was taken up by the same set of investigators. The first part of the sample was utilised to obtain preliminary estimates of certain major items quickly as the National Income Committee wanted the estimates by April 1951.

#### FIELD ORGANISATION

3.5. The staff employed for the survey under the Director of Industrial Statistics consisted of (i) an Officer on Special Duty, (ii) six Regional Research Officers and (iii) thirty-two Investigators. India was divided into seven regions for the field work. The work in Assam region was placed under the Statistics Authority, Assam. The remaining six regions were under the six Regional Research Officers. Each region was further divided into a number of investigator areas.

3.6. The field work began in the middle of January 1951. The establishments included in the first part of the sample were surveyed by the third week of March. The survey of the establishments of the second part was then taken up and was completed by the middle of June 1951.

#### SCRUTINY AND ANALYSIS OF DATA

3.7. It was arranged that the Indian Statistical Institute would analyse the data and that before sending the completed schedules to the Institute, the Office of the Director of Industrial Statistics would scrutinise the returns.

3.8. The schedules completed by the Investigators were forwarded to the Head Office at Simla after scrutiny by the Regional Research Officers. These were further scrutinised by the Officer on Special Duty and the Director of Industrial Statistics and, where necessary, the returns were referred back to the Regional Research Officers for correction of errors or omissions noticed. The returns were then sent to the Indian Statistical Institute for analysis.

3.9. Although it was arranged at first that the work of analysis should start straightaway without further scrutiny, some checking was, however, found necessary in the tabulation stage when some minor defects such as disagreement between the components and the sub-totals, ambiguous entries etc., were discovered.

3.10. In addition, the scrutiny of the tabulated results also constituted an important part of the analysis work; where the tabulation consisted of building up roughly 45,000 estimates some broad and suitable criteria for checking individual estimates had to be adopted. The estimates were, therefore, studied in the light of a number of criteria some of which were (i) ratio of fixed to working capital; (ii) per-hour earning of a worker in different establishments of the same industry and (iii) ratios between the excess of value of output over input and labour charges on the one hand and the total capital employed on the other.



3.11. For both the years 1949 and 1950 the ratios were worked out for each industry. From the central tendency and scatter of these ratios the doubtful cases were noticed easily and both the computation sheets as well as the completed schedules were scrutinised again.

3.12. Although 1746 establishments were surveyed, the total schedules for analysis stood at 1742 after scrutiny and rejections. For each establishment, the number of items of information collected was 37 for each year, that is 74 in total for the two years. The estimates in respect of these 74 characters were obtained for each of the 589 strata. Besides, summary estimates in respect of the industries and of the different States were made for each of the 74 characters.

3.13. As already stated, the establishments in the first part of the sample were surveyed by the third week of March and the completed schedules after scrutiny were sent to the Indian Statistical Institute by the 31st of March 1951. The main results based on this part were furnished to the National Income Committee by the first week of April 1951 as required by them. The figures of both the parts were taken up for analysis when the whole survey was over. The tabulation of all the details was completed and the tables were passed on to the National Income Committee by the end of February 1952.

#### Cost

3.14. The budget estimate of the cost of this sample survey was just below a lakh of rupees. Round figures of the actual cost under different broad headings are given below.

Planning	Rs. 10,000
Field work	Rs. 80,000
Processing and analysis	Rs. 25,000
	<hr/>
	Rs. 1,15,000

3.15. It should be noted that in the budget estimates the cost of processing, analysis etc., was estimated at only Rs. 3,000 and this was a clear underestimate. The Indian Statistical Institute undertook to do the analysis in any case without regard to the cost which amounted to Rs. 25,000 approximately. Thus, although the budgeted figure was roughly Rs. 93,000, the actual cost was Rs. 1,15,000.



## CHAPTER FOUR

## RELIABILITY OF ESTIMATES

4.1. There are scarcely any data available in published form which can be used to test the reliability of the results of the survey in an exact way. For any proper comparison the coverage of the figures must be the same. Results of 1949 and 1950 Census of Manufactures have since been published by the Directorate of Industrial Statistics. The census was restricted to 29 groups of industries. The number of factories covered was 6758 and 7099 in 1949 and 1950 respectively. The coverage of the sample survey was 7928 factories in both the years, so far as these 29 groups of industries were concerned. Because of this reason of wider coverage of the sample survey, even the estimates for 29 groups of industries are not strictly comparable with the census results. But this factor should make the census results lower than the sample estimates. In Table (4.1) comparative figures are indicated in respect of eight important items of information for the 29 groups of industries covered by the census.

TABLE (4.1): COMPARISON BETWEEN SAMPLE SURVEY AND CENSUS RESULTS:  
1949 AND 1950

item	unit	1949			1950		
		sample survey	census	per cent difference	sample survey	census	per cent difference
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. fixed capital	Rs. crore	316	228	38.6	346	258	34.1
2. working capital	„	386	282	37.0	375	356	5.0
3. total invested capital	„	702	510	37.6	721	614	17.4
4. emoluments of labour	„	197	177	11.1	184	172	7.1
5. value of input	„	803	687	15.7	833	726	13.4
6. value of output	„	1130	976	15.8	1164	1028	13.2
7. 'workers' per day	lakh	17	15	13.3	16	15	6.7
8. man-hours	crore	1122	969	15.8	1156	1023	13.0
9. factories covered	number	7928	6250	26.8	7928	6605	20.0
10. sample size	„	1013	—	—	1013	—	—

4.2. It will be seen from the above table that in all cases the sample estimates are greater than the census results and that the gap between the results is smaller in 1950 than in 1949. The factory coverage of the census in 1950 was much larger than in 1949, but still smaller than the survey. The divergence thus appears to decrease with the decrease in the gap between factory coverage of the census and the sample survey. The survey results are reasonably higher than the census results. There is, however, rather a large discrepancy regarding fixed capital. It is a difficult field



for collection of information, whichever be the method of collection. No definite observation on the merits of either of the results is possible, unless a factory-to-factory comparison, at least in the case of the sample factories, is made. But such details of the census data are not available with the NSS.

4.3. As mentioned earlier, the field inquiry was done in two parts. In the first part 798 sample establishments from 61 industries including their sub-groups were covered and in the second part 944 establishments were covered from 68 industries including the sub-groups. Because the samples in the two parts were not always scattered over common strata or even industries, the two half samples were not strictly comparable sub-samples. Hence, the estimated results got from the two parts of the sample are not comparable either. Since, however, due to time programme of analysis, the results of the two parts are available separately, a comparison between the two sets of figures may be of some interest. The figures are given in Table (4.2).

TABLE (4.2) : ESTIMATES OF SELECTED ITEMS AS OBTAINED FROM THE TWO PARTS OF THE SAMPLE

item	sample estimates : Rs. (crore)						per cent difference	
	first part		second part		combined		1949	1950
	1949	1950	1949	1950	1949	1950		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. fixed capital	530.88	566.76	544.02	596.28	538.00	582.76	— 2.44	— 5.07
2. working capital	577.91	587.77	487.90	459.19	530.05	518.20	16.98	24.81
3. amount received by labour	267.78	263.91	261.91	244.63	264.60	253.36	2.22	7.61
4. value of input	1149.59	1195.12	1007.90	1072.63	1072.82	3128.69	12.16	10.85
5. value of output	1777.21	1815.58	1398.25	1495.60	1571.39	1642.18	24.12	19.49
6. sample size	798	798	944	944	1742	1742		

4.4. It will be seen that in 3 of the 5 cases the agreement is not good. But it is not possible to analyse the reason of such disagreements because the two parts are not strictly comparable.

#### CONCLUDING REMARKS

4.5. The extent to which the respondents gave correct information is not known. The method of collection of data was that the investigator would visit the establishments or the owners and complete the schedules there. As the complete addresses of the sample units were available, there were no difficulties in locating the establishments. The main difficulties in obtaining data were, in the first place, that the owners did not maintain statistical records of production and employment in the way these were wanted in the questionnaire and secondly, that in some cases



the owners were unwilling to furnish these particulars to a Government agency because they were suspicious of the motives of the survey. Where the records were incomplete the investigators obtained the estimates from the owners of the establishments. In regard to the second type of difficulty the investigators had to explain the purpose of the survey to the owners and to impress on them that these returns were not meant for income-tax purposes. Of the 1,885 samples selected 1,746 were actually surveyed. The proportion not surveyed was 7.4%.

4.6. It may be noted that this problem of non-response and deliberate furnishing of inaccurate data is not a problem limited to sample surveys only but also common to censuses. The actual method of collecting data in the Censuses of Manufactures is by mailing questionnaires with the system of the field staff of the State Statistics Authorities assisting the occupiers of factories in filling the returns completely and accurately where necessary. The method followed in the sample survey was that the investigator had to visit the selected establishments in every case with a view to minimising non-response.

4.7. In addition, as the sample size was only a fraction of the total establishments in the country, the number of schedules completed during the survey was considerably small. These completed schedules could, therefore, be scrutinised by the Regional Research Officers and then by the Officer on Special Duty and ultimately, in some cases, by the Director of Industrial Statistics. Wherever it was found necessary, the schedules were referred back to the investigators so that correct data might be obtained after clearing up the inconsistencies with the owners of the establishments. Thus it is reasonable to say that the data obtained are reliable.

## CHAPTER FIVE

## SUMMARY RESULTS

5.1. The tabulated results give the estimates for 37 items for the years 1949 and 1950 arranged by industries. A summary of some of the results is given in this chapter. It may be mentioned in passing that the figures relate entirely to manufacturing industries and hence exclude other branches of productive activity such as trade, transport, commerce, mining etc. They also exclude particulars of small scale manufacturing establishments, not covered by Section 2(j) of the Factories Act of 1934.

5.2. The All-India estimates of the value of fixed and working capital, rent paid by establishments, amount received by labour, value of raw materials, value of input, value of output and the difference between the values of input and output are given in Table (5.1). The value of fixed capital, that is, the value of land and buildings, plant and machinery and other fixed assets amounted to Rs. 538 crore in 1949 and Rs. 583 crore in 1950. The values were based on the original costs of the fixed capital plus the cost of improvements made less the amount written off as discarded. The rents paid for using fixed capital on lease amounted to less than 1 per cent of the value of fixed capital owned by the establishments. The totals of fixed and working capital employed by the manufacturing industries were

TABLE (5.1): ESTIMATES OF VALUE OF SOME SELECTED ITEMS  
RELATING TO MANUFACTURING INDUSTRIES OF  
INDIA IN 1949 AND 1950

item	Rs. (crore)	
	1949	1950
(1)	(2)	(3)
1. fixed capital	538.00	582.76
2. working capital	530.05	518.20
3. rent	3.55	3.60
4. amount received by labour	264.60	253.36
5. value of raw materials	1014.23	1067.32
6. value of input	1072.82	1128.69
7. value of output	1571.39	1642.18
8. difference (7-6)	498.57	513.49
9. sample size	1742	1742

Rs. 1068 crore in 1949 and Rs. 1101 crore in 1950. The amounts received by labour including workers and other employees amounted to Rs. 265 crore in 1949 and Rs. 253 crore in 1950. The values of raw materials used were Rs. 1014



crore and Rs. 1067 crore in 1949 and 1950 respectively. The difference between the values of output and input, that is, the value added by manufacture gross of depreciation amounted to Rs. 499 crore in 1949 and Rs. 513 crore in 1950.

TABLE (5.2) : A FEW SELECTED ITEMS RELATING TO MANUFACTURING INDUSTRIES IN 1949 AND 1950

item	1949 (crore)	1950 (crore)
(1)	(2)	(3)
1. number of working days	0.34	0.35
2. total number of workers employed per day	0.24	0.23
3. total number of persons other than worker employed per day	0.03	0.03
4. total labour employed per day	0.27	0.26
5. man-hours worked by workers	519.66	493.36
6. electricity consumed (kwh)	198.20	202.73
7. sample size	1742	1742

5.3. The total number of working days of all manufacturing establishments was estimated at 0.34 crore for 1949 and 0.35 crore for 1950. The total number of workers employed per day was estimated at 0.24 crore in 1949 and 0.23 crore in 1950. When the employees other than the workers are taken into consideration the total of labour employed amounted to 0.27 crore in 1949 and 0.26 crore in 1950. The total quantity of electricity in kwh consumed by the manufacturing establishments was estimated to be 198 crore and 203 crore in 1949 and 1950 respectively.

#### PARTICULARS BY INDUSTRY-GROUPS

5.4. The particulars for six most important groups of industries in India judging from their value of output are given below as a matter of interest. The figures for the two years are shown separately in Table (5.3).

5.5. It will be seen that the six industry groups in order of their importance are (1) manufacture of textiles, (2) manufacture of food and beverage, (3) manufacture of chemicals and chemical products, (4) manufacture of basic metals, (5) ginning, pressing, decorticating and similar services to agricultural products, and (6) manufacture of machinery excluding electrical machinery and appliances. The lighter industries have thus much predominance in the pattern of our manufacturing activities.



# NATIONAL SAMPLE SURVEY : MANUFACTURING INDUSTRIES, 1949 AND 1950

TABLE (5.3) : ESTIMATES OF SELECTED ITEMS FOR SOME INDUSTRY GROUPS IN  
1949 AND 1950

industry group	number of sample units	fixed capital Rs. (crore)	working capital Rs. (crore)	number of workers (thousand)	value of input Rs. (crore)	value of output Rs. (crore)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
<b>1949</b>						
1. basic metals	51	51.89	42.93	93	47.86	82.48
2. chemicals and chemical products	135	50.93	48.37	1,17	183.95	211.08
3. food and beverage	243	87.52	75.47	2,65	190.26	290.46
4. ginning, pressing and similar services to agricultural products	160	26.77	7.43	1,38	61.31	64.91
5. machinery excluding electrical machinery	147	31.11	28.98	1,48	29.28	57.55
6. textile	521	123.37	205.81	11,71	388.11	577.20
7. total (1 to 6)	1,257	371.59	408.99	19,32	900.77	1283.68
8. total of all industries	1,742	538.00	530.05	24,24	1072.82	1571.39
<b>1950</b>						
1. basic metals	51	56.10	43.60	91	55.64	90.14
2. chemicals and chemical products	135	52.31	50.37	1,06	181.79	215.44
3. food and beverage	243	96.40	70.30	2,71	201.63	317.43
4. ginning, pressing and similar services to agricultural products	160	26.55	5.88	1,41	61.53	66.47
5. machinery excluding electrical machinery	147	35.43	28.85	1,55	32.41	62.76
6. textile	521	134.44	200.80	10,85	398.92	570.87
7. total (1 to 6)	1,257	401.23	399.80	18,49	931.92	1323.11
8. total of all industries	1,742	582.76	518.20	23,37	1128.69	1642.18

5.6. Table (5.4) shows the value of fixed capital per employed worker and the value of output per employed worker in the six groups of industries. The figures for all the industries grouped together are also given. The value of fixed capital per worker was highest in the basic metal industries. Next in order of ranking the groups are : chemical and chemical products industries, food and beverage industries, machine manufacturing industries, ginning, pressing and similar industries, and textile industries. It may be noted that this ratio for all industries was higher than the ratio for the first six groups of industries taken together.

5.7. The value of output per worker was, however, highest in the chemical and chemical products industries. In order of ranking, the other industries are food and beverage, basic metals, textile, ginning, pressing and similar servicing, and



lastly manufacture of machinery. The productivity of workers of all industries taken together was roughly of the same order as that of the first six industry groups.

TABLE (5.4): ESTIMATES OF FIXED CAPITAL AND OUTPUT PER WORKER FOR SOME INDUSTRY GROUPS IN 1949 AND 1950

industry group	estimates per worker (Rs.)			
	fixed capital		output	
	1949	1950	1949	1950
(1)	(2)	(3)	(4)	(5)
1. basic metals	5,580	6,165	8,869	9,905
2. chemicals and chemical products	4,353	4,935	18,041	20,325
3. ginning, pressing and similar services to agricultural products	1,940	1,883	4,704	4,714
4. food and beverage	3,303	3,557	10,961	11,713
5. machinery excluding electrical machinery	2,102	2,286	3,889	4,049
6. textile	1,054	1,239	4,929	5,261
7. total (1 to 6)	1,923	2,170	6,644	7,156
8. total of all industries	2,219	2,494	6,483	7,027

5.8. When compared between the two years, the value of fixed capital per worker increased to some extent from 1949 to 1950 in all the groups except in ginning, pressing and similar industries. The value of output per worker also increased in varying extent from 1949 to 1950. Without going into the detailed tables of individual industries it is difficult to indicate as to how much of this increase was due to price variation and how much due to increase in quantity.

#### PARTICULARS BY TEN MAJOR INDUSTRIES

5.9. The sample survey of manufacturing industries, as stated earlier, covered 61 industries. The total number of establishments for which estimates have been made was 17,377. Out of these, factories belonging to the ten major manufacturing industries number 3050. Their distribution along with the samples taken in each case is as in Table (5.5).

5.10. The ten major industries for selective review are cotton textile, jute, iron and steel, tea, sugar, chemicals, paper and paper board, tobacco, cement, and paints and varnishes. The number of factories covered by these ten industries was about 18 per cent of the total number of factories in all the industries, but accounted for 55.14 and 54.68 per cent of the total invested capital in all industries in 1949 and

NATIONAL SAMPLE SURVEY : MANUFACTURING INDUSTRIES, 1949 AND 1950

TABLE (5.5) : TOTAL NUMBER OF FACTORIES AND THE NUMBER OF  
SAMPLE FACTORIES (RELATING TO TEN MAJOR MANU-  
FACTURING INDUSTRIES) IN 1949 AND 1950

industry		total number of factories	number of sample factories
(1)		(2)	(3)
1.	cement	17	11
2.	heavy chemicals	210	27
3.	cotton textile	755	302
4.	iron and steel	203	22
5.	jute textile	104	100
6.	paper and paper board	52	15
7.	paints and varnishes	48	8
8.	sugar	431	89
9.	tea	1080	78
10.	tobacco	150	19
11.	total (1 to 10)	3050	671
12.	total of all industries	17,377	1742

1950 respectively. Table (5.6) sets out the position of these major industries of India with regard to their size as measured by capital outlay and the extent to which they own fixed assets in comparison with the position of all industries.

5.11. It will be seen that cotton textile, iron and steel, and jute by themselves make for about 34 per cent of the total capital outlay in all industries. An observation of the figures of fixed capital for the ten industries, in the following table brings out an upward trend in the fixed capital investment in these industries in 1950 over 1949. This is a sign of development of these industries. The figures of working capital for these major industries, however, indicate in 1950 a falling tendency as compared to those of 1949, the only notable exception being tea industry where working capital rose in 1950 by 19 per cent over that in the previous year. In so far as the rented fixed assets are concerned the year 1950 appears to have been marked with an effort on the part of these industries to reduce rent payments on fixed assets—from Rs. 58 lakh in 1949 to Rs. 51 lakh in 1950—by owning more fixed assets. The amount of rent paid on fixed assets for all industries recorded a slight increase from Rs. 3.55 crore in 1949 to Rs. 3.60 crore in 1950. Exceptions in this regard are paper and paper board, tea and cement industries, even though the first two of these otherwise effected an increase in their working capital.



TABLE (5.6) : ESTIMATES OF FIXED AND WORKING CAPITAL AND RENT PAID ON FIXED ASSETS (IN TEN MAJOR MANUFACTURING INDUSTRIES) IN 1949 AND 1950

Rs. (crore)				
industry	fixed capital	worki capital	total invested capital	rent paid on fixed assets
(1)	(2)	(3)	(4)	(5)
<b>1949</b>				
1. cement	10.74	5.91	16.65	0.03
2. chemicals	18.14	12.18	30.32	0.08
3. cotton textiles	79.53	150.82	230.35	0.11
4. iron and steel	43.67	24.20	67.87	0.02
5. jute textiles	26.74	38.93	65.67	0.09
6. paints and varnishes	0.72	1.98	2.70	0.01
7. paper and paper board	15.08	6.39	21.47	0.00
8. sugar	20.66	37.55	58.21	0.11
9. tea	50.80	23.02	73.82	0.09
10. tobacco	6.40	15.44	21.84	0.04
11. total (1 to 10)	272.48	316.42	588.90	0.58
12. percentage to total of all industries	50.65	59.70	55.14	16.34
13. total of all industries	538.00	530.05	1068.05	3.55
<b>1950</b>				
1. cement	10.70	5.53	16.23	0.04
2. chemicals	17.91	12.17	30.08	0.08
3. cotton textiles	88.30	146.96	235.26	0.10
4. iron and steel	47.75	23.34	71.09	0.02
5. jute textiles	28.62	37.53	66.15	0.05
6. paints and varnishes	0.55	1.61	2.16	0.01
7. paper and paper board	17.93	6.24	24.17	0.00
8. sugar	26.69	30.59	57.28	0.09
9. tea	53.00	27.40	80.40	0.09
10. tobacco	6.83	12.39	19.22	0.03
11. total (1 to 10)	298.28	303.76	602.04	0.51
12. percentage to total of all industries	51.18	58.62	54.68	14.17
13. total of all industries	582.76	518.20	1100.96	3.60



# NATIONAL SAMPLE SURVEY : MANUFACTURING INDUSTRIES, 1949 AND 1950

5.12. The gross income of 61 industries was Rs. 1571 crore in the year 1949 and Rs. 1642 crore in the year 1950. They are distributed in Table (5.7).

TABLE (5.7) : PRODUCTION ACCOUNT OF 61 MANUFACTURING INDUSTRIES  
IN 1949 AND 1950

		Rs. (crore)	
item	1949	1950	
(1)	(2)	(3)	
<b>A. Value of production</b>			
1. products and by-products	1529.48	1601.95	
2. work done for other concerns	41.91	40.23	
3. total (1 + 2)	1571.39	1642.18	
<b>B. Value of input</b>			
4. raw materials and chemicals etc.	1014.21	1067.16	
5. fuels, lubricants etc.	51.31	53.63	
6. work done by other concerns	7.30	7.90	
7. total (4 to 6)	1072.82	1128.69	
C. Depreciation estimated @ 7%	37.66	40.79	
<b>D. Value added by manufacture (net of depreciation)</b>			
8. salaries, wages and other benefits received by labour	264.60	253.36	
9. balance available for other purposes	196.31	219.34	
10. total (8 + 9)	460.91	472.70	

5.13. The figures of depreciation on fixed assets were not collected separately in this survey. However, the Income Tax Manual, Part II, 1954, in accordance with Section 10(2)(vi) of the Income Tax Act, 1922, prescribes a general rate of depreciation on fixed assets at 7 per cent of the value of such assets. Worked on this basis the amount of depreciation on fixed assets comes to Rs. 37.66 crore in 1949 and Rs. 40.79 crore in 1950. Thus, the value added by manufacture, net of depreciation, comes to Rs. 460.91 crore in 1949 of which Rs. 264.60 crore or about 57.4 per cent were shared by wages and salary earners. In the following year (1950), the value added, net of depreciation, comes to Rs. 472.70 crore of which Rs. 253.36 crore, or about 53.6 per cent went to wages and salary earners.

5.14. The comparative values in the ten major industries as against those in all the industries, in respect of the components of the gross income and gross expenditure are examined later in details under the separate section devoted to these items. The comparative gross income in the ten major industries was Rs. 860 crore in 1949, and Rs. 883 crore in 1950, forming respectively 54.7 and 53.8 per cent of the gross income of all industries.



## INPUT

5.15. *Cost of materials* : The questionnaire called for data on the quantity and purchase value of each material consumed during the year. Only materials, which were purchased, have been included. Materials made in the factory have not been included. The purchase value of the quantity of material purchased during the year has been taken as equal to the cost of material landed at the factory, *i.e.*, any expense incurred in transporting the materials to the factory have been added to the payment made to the seller of the material unless transport was carried out by the factory's own staff. If any duty was paid by the factory, it has also been added to the amount paid to the seller. Particulars relating to goods, which were not subjected to any manufacturing process but were merely bought and re-sold in the same condition as received, have been excluded. The total amount paid to other firms or factories for work done on materials given out to them plus transport and any other charges incurred on these goods have been included.

5.16. *Fuel and electric energy used* : The quantities of the several kinds of fuel (coal, coke, fuel oil and gas) used, the quantity of electric energy purchased and the quantity of water used by manufacturing establishments have been reported together with the cost of each. Fuel, electricity etc., produced in the factory have not been included. If any electricity generated (or coal gas produced) is sold to any person or transferred to allied concerns, such electricity (or coal gas) has been regarded as a product.

5.17. The All-India figures of costs in all industries are shown in Table (5.8). It would appear that the total costs in all industries recorded an increase of Rs. 56

TABLE (5.8) : ESTIMATES OF COST ITEMS

	Rs. (crore)	
items of cost	1949	1950
(1)	(2)	(3)
1. raw materials	1014.21	1067.16
2. fuels etc.	51.31	53.63
3. work done for the factory by others	7.30	7.90
4. total	1072.82	1128.69

crore or 5.2% in 1950 over the previous year. Of the total cost the cost of raw materials formed on an average 94.5 per cent.



# NATIONAL SAMPLE SURVEY : MANUFACTURING INDUSTRIES, 1949 AND 1950

5.18. The increase or decrease in the costs of materials etc., for each of the ten major industries is shown in Table (5.9).

TABLE (5.9) : INCREASE OR DECREASE IN COSTS OF MATERIALS ETC.  
IN 1950 OVER 1949

Rs. (lakh)

industry	raw materials and chemicals	fuel, lubricants and electricity	work done for the factory by other concerns	total input
(1)	(2)	(3)	(4)	(5)
1. cement	+ 141.28	+ 22.79	+ 0.44	+ 164.51
2. chemicals	+ 187.19	+ 24.29	- 0.25	+ 211.23
3. cotton textile	+1345.53	+ 26.71	- 25.19	+1347.05
4. iron and steel	+ 89.67	+ 29.23	- 1.64	+ 117.27
5. jute textile	- 775.13	- 21.91	+ 4.03	- 793.02
6. paints and varnishes	- 42.51	- 0.10	-	- 42.60
7. paper and paper board	+ 98.18	+ 39.23	- 1.34	+ 136.07
8. sugar	+ 98.90	- 9.46	+ 0.87	+ 90.31
9. tea	+ 311.38	+ 24.47	+ 7.88	+ 343.73
10. tobacco	+ 386.19	+ 0.01	- 1.39	+ 384.81
11. total	+1840.68	+135.26	- 16.59	+1959.36

5.19. For the ten major industries these costs as compared to those for all the industries were slightly lower at 49.27 per cent in 1950 compared to 50.01 per cent in 1949. Industry-wise, the raw materials and chemicals consumed were on the increase in 1950 compared to the previous year, paints and varnishes and jute textiles being exceptions.

5.20. The fuel cost of the ten major industries in 1949 and 1950 as a percentage of corresponding figure for all the industries, however, remained more or less stationary at 60 per cent. Here again in sugar, paints and varnishes, and jute textile industries, the fuel cost in the latter year was lower. The other seven industries exhibited an increase in their fuel costs.

5.21. The cost of work done by other concerns for 10 industries compared to the corresponding item for all industries declined from 42 per cent in 1949 to 37 per cent in 1950. In cotton textile industry, in particular, where the work done for the industry by other concerns generally covers quite a few processes, the cost on this account was lower by Rs. 25.19 lakh in the latter year, indicating thereby the industry's capacity to complete many such processes by itself. Paper and paper board, iron and steel, and chemicals were other industries in order of precedence which effected economy in 1950 over the year 1949 in respect of payments for work done for them by other concerns. In the jute and tea industries, on the other hand, the cost of work done for them by other concerns showed an increase in the latter year.



### PRODUCTS AND BY-PRODUCTS

5.22. The value of products and by-products in all industries was Rs. 1529.48 and Rs. 1601.95 crore in 1949 and 1950 respectively which meant an increase in the latter year by 4.7 per cent. On the other hand, the value of work done for others in all industries was 4 per cent lower in 1950 than that in the previous year. This is shown in Table (5.10).

TABLE (5.10) : ESTIMATES OF OUTPUT FOR ALL INDUSTRIES IN 1949 AND 1950

item	Rs. (crore)	
	1949	1950
(1)	(2)	(3)
1. products and by-products	1529.48	1601.95
2. work done for others	41.91	40.23
3. total	1571.39	1642.18

5.23. As a percentage of all industries' total, the value of products and by-products for the ten major industries in the year 1950 registered a nominal decline over the year 1949 although in absolute terms these items for the ten major industries showed an increase of 2.7 per cent in 1950 over the previous year. The following table presents the percentage increase or decrease in the value of products and by-products as well as that in the value of work done by the factory for others in 1950 over the year 1949.

TABLE (5.11) : PERCENTAGE INCREASE OR DECREASE IN THE VALUE OF OUTPUT IN 1950 OVER 1949—TEN MAJOR INDUSTRIES

industry	products and by-products	work done for others
(1)	(2)	(3)
1. cement	+37.14	+23.45
2. chemicals	+18.53	+30.20
3. cotton textile	-4.00	+8.84
4. iron and steel	-1.15	+20.83
5. jute textile	+2.41	+28.51
6. paints and varnishes	-22.54	—
7. paper and paper board	+18.23	—
8. sugar	+6.41	-45.22
9. tea	+16.25	-59.73
10. tobacco	+11.65	-18.31



5.24. It is evident from the above that the value of products and by-products was on an increase in the year 1950, in order of precedence, in the cement industry, chemicals, paper and paper board, tea, tobacco, sugar and jute textiles while it was lower in respect of paints and varnishes, cotton textiles and iron and steel industries.

5.25. The work done by factory for customers is a source of revenue to the industry. The value of such work done in the ten major industries as a percentage of this work for all industries recorded an increase of 0.3 per cent in 1950 over 1949. For these ten industries themselves, the year 1950 recorded a rise in value in this sphere of the order of 9.2 per cent over that in the year 1949. The percentage increase in value in the year 1950 compared to the previous year was 30.20 in chemicals, 28.51 in jute, 23.45 in cement, 20.83 in iron and steel and 8.84 in cotton; the percentage decrease in value was of the order of 59.73 in tea, 45.22 in sugar, 18.31 in tobacco.

#### VALUE ADDED BY MANUFACTURE

5.26. In our survey the concept of 'net value of production' has not been used. We have used a similar concept, 'value added by manufacture'. This measures the increase in the total value of commodities by the manufacturing process and is calculated by subtracting the cost of materials, supplies, containers, fuel, purchased electric energy, contract work, and the depreciation of fixed assets from the total value of products and work done by the industry for customers.

5.27. The figure thus calculated is somewhat larger than the net value of production because many miscellaneous expenses, such as commission on sales, insurance and advertising, have not been taken into account. Therefore, it must not be inferred that when wages and salaries and undistributed profits are deducted from these values added by manufacture the whole of the residue is available for non-wage factor payments. The relevant figures are given in Table (5.12).

TABLE (5.12): VALUE ADDED BY MANUFACTURE IN 1949 AND 1950  
—ALL INDUSTRIES

item	Rs. (crore)	
	1949	1950
	(2)	(3)
(1)		
1. value of input	1072.82	1128.69
2. depreciation estimated @7%	37.66	40.79
3. value of output	1571.39	1642.18
4. value added (net of depreciation)	460.91	472.70

5.28. The value added by manufacture in all industries stood at Rs. 460.91 and Rs. 472.70 crore in the year 1949 and 1950 respectively, marking an increase of 2.56 per cent in the latter year. The corresponding figures for ten major industries are given in Table (5.13).



TABLE (5.13) : VALUE ADDED BY MANUFACTURE IN 1949  
AND 1950—TEN MAJOR INDUSTRIES

item	Rs. (crore)	
	1949	1950
	(2)	(3)
1. value of input	536.56	556.15
2. depreciation estimated @7%	19.07	20.88
3. value of output	860.15	883.33
4. value added (net of depreciation)	304.52	306.30

5.29. The table above discloses that the value added by manufacture in the ten major industries in comparison with that for all the industries recorded a decline of 1.3 per cent in 1950 over the previous year. The comparative total value added by manufacture for these ten industries alone, however, showed a rise of 0.58 per cent in 1950 over the year 1949.

5.30. It would be of interest to have an idea of the value added per worker in all industries as well as in the ten major industries in 1949 and 1950. The value added per worker for all industries was Rs. 1901 in 1949 and Rs. 2023 in 1950 as against Rs. 2108 and Rs. 2247 respectively for the ten industries.

#### INVESTMENT

5.31. *Capital structure* : All particulars under this head are as they were on the 31st December 1949 or 1950 or the date on which the factory last closed accounts. 'Value' in all the headings specified under the item 'productive capital employed' should be taken to mean value according to the books of the factory on the date to which the particulars furnished under this item relate. For items of fixed capital, these are the original cost plus the cost of improvements made less amount written off. In case a factory occupies only a portion of any building or any piece of land, particulars relating to only that portion had been included. In the case of any item of fixed capital which had been leased or rented, the rent paid had been shown separately. In calculating this rent any lump sum consideration that was originally paid for securing the items of fixed capital in question either on lease or on rent, the present book value of the amount originally paid had been included in the amount of the rent.

5.32. The invested capital in all industries stood at Rs. 1068.05 and Rs. 1100.96 crore in 1949 and 1950 respectively showing a rise of about 3 per cent in the latter year. Compared to this, the fixed capital in all industries was Rs. 538.00 and Rs. 582.76 crore during 1949 and 1950, recording a rise of 8.32 per cent in the latter year. The comparative figures for working capital were Rs. 530.05 and

# NATIONAL SAMPLE SURVEY : MANUFACTURING INDUSTRIES, 1949 AND 1950

Rs. 518.20 crore during 1949 and 1950 respectively showing a decline of 2.24 per cent in the latter year. This is brought out in Table (5.14).

TABLE (5.14) : ESTIMATES OF FIXED AND WORKING CAPITAL  
FOR ALL INDUSTRIES IN 1949 AND 1950

		Rs. (crore)	
item		1949	1950
(1)		(2)	(3)
1.	fixed capital	538.00	582.76
2.	working capital	530.05	518.20
3.	total	1068.05	1100.96

5.33. The total capital investment in the ten major industries comprised, as already observed, 55.14 and 54.68 per cent of the capital invested in all the industries taken together during the years 1949 and 1950 respectively. On the other hand, the fixed capital invested in these ten major industries comprised respectively 50.65 and 51.18 per cent during 1949 and 1950 of the total fixed capital outlay in all the industries. The working capital invested in these industries formed 59.70 and 58.62 per cent during the year 1949 and 1950 respectively of the total working capital in all the industries. While it would be needless to repeat here the table giving figures of fixed, working and invested capital, for the ten major industries, it would be no doubt instructive to observe the relationship between working capital and total capital invested in Table (5.15).

TABLE (5.15) : WORKING CAPITAL AS PERCENTAGE OF  
INVESTED CAPITAL IN 1949 AND 1950—TEN  
MAJOR INDUSTRIES

industry		1949	1950
(1)		(2)	(3)
1.	cement	35.52	34.07
2.	chemicals	40.16	40.44
3.	cotton textile	65.48	62.47
4.	iron and steel	35.65	32.83
5.	jute textile	59.29	56.73
6.	paints and varnishes	73.44	74.72
7.	paper and paper board	29.74	25.82
8.	sugar	64.51	53.40
9.	tea	31.19	29.04
10.	tobacco	70.70	64.47
11.	all industries	49.63	47.07



5.34. The year 1950 is marked by a general tendency in all these industries for the working capital to be lower as compared to the previous year. During the year 1949, the proportion of the working capital to invested capital was the highest at 73.44 per cent in the paints and varnishes industry followed by 70.70 per cent in tobacco, 65.48 per cent in cotton textile, 64.51 per cent in sugar and 59.29 per cent in jute textile industry.

5.35. The output as a percentage of the invested capital in all industries was of the order of 147 in 1949 and 149 in 1950. As against this the value added for all industries as per cent of the invested capital was broadly 43 per cent of the invested capital for both the years. Table (5.16) shows the output and value added as percentages of invested capital.

TABLE (5.16) : GROSS AND NET RATIOS OF OUTPUT TO INVESTED CAPITAL  
IN 1949 AND 1950

industry	percentage of invested capital			
	output		value added*	
	1949	1950	1949	1950
(1)	(2)	(3)	(4)	(5)
1. cement	74.28	104.50	28.78	47.71
2. chemicals	75.57	90.33	34.98	42.42
3. cotton textile	165.56	155.63	56.86	43.23
4. iron and steel	85.51	80.78	37.71	33.10
5. jute textile	220.32	224.00	50.71	67.40
6. paints and varnishes	166.83	161.60	58.55	46.45
7. paper and paper board	63.75	66.97	20.62	22.91
8. sugar	163.37	176.62	50.32	59.42
9. tea	124.91	133.32	73.42	81.58
10. tobacco	161.69	205.08	44.61	51.90
11. total (1 to 10)	146.06	146.72	51.71	50.88
12. total of all industries	147.14	149.12	43.15	42.93

5.36. The value added as a percentage of the invested capital is markedly higher in the ten major industries. Among these industries the highest output expressed as a percentage of invested capital is recorded in order of precedence in jute, tobacco, and sugar industries whereas the value added as percentage of the invested capital is the highest in tea, jute and sugar industries.

#### LABOUR AND THEIR EARNINGS

5.37. According to the Census of 1951 the total population of India was 3613 lakh for the year 1951. The Census Report observes that there is a recurring net annual increase in our population of 1.3 per cent or 44 lakh. Applying this

\* net of depreciation



annual rate of increase to the population figure for 1951, we arrive at a population figure of 3525 lakh for the year 1949, the corresponding figure being 3569 lakh for the year 1950. Table (5.17) gives the figures of total population, aggregate self-supporting working population, and the self-supporting working population in industries.

TABLE (5.17) : TOTAL AND WORKING POPULATION COMPARED TO THE WORKING POPULATION IN INDUSTRIES\*

year	total population	working population			
		total	industries	manufacturing industries	ten major industries
(1)	(2)	(3)	(4)	(5)	(6)
1949	3,525	1,009	89	27	16
1950	3,569	1,021	91	26	15

5.38. The 1951 Census enumerated the total working population in the country at 1044 lakh of persons or 28.62 per cent of the total population. Out of this, 334 lakh persons or 9.24 per cent were reported to be engaged in non-agricultural occupations. The workers engaged in industries of all types and sizes (*i.e.*, in processing and manufacturing) including such establishments as are covered by the Factories Act were, however, returned at 92 lakh in 1951 or 2.54 per cent of the total population. Applying the percentage of working population to total population in 1951 to the total population figures for 1949 and 1950, we arrive at the figures of total working population for these years as shown in the table above. Similarly, the workers engaged in industries during the years 1949 and 1950 have been arrived at by applying the percentage of workers engaged in industries in 1951, *i.e.*, 2.54 per cent to the total population figures of the years 1949 and 1950.

5.39. Taking the proportion of self-supporting working population to total population and that of the self-supporting working population in industries to the total population as obtained in the Census figures for 1951, the working population works out to be 1009 lakh in 1949 and 1021 lakh in 1950. The working population in industries, on the same basis of calculation, comes to 89 lakh in 1949 and 91 lakh in 1950. These figures are broadly comparable to the figures of persons employed in all industries, as well as in the ten major industries, as estimated in our survey. The figures of total working population in the 61 organised industries as given by our survey were 27 lakh or 30 per cent in 1949 and 26 lakh or 29 per cent in 1950. The working population in the ten major industries was 16 lakh in 1949 and 15 lakh in 1950. When compared to the working population in organised industries alone, the figures of persons employed in the ten major industries comprised 58 per cent and 57 per cent in 1949 and 1950 respectively. Coming to individual industries we find that iron and steel, cotton textile and jute between them employed 45 per cent of the total workers in all the industries.

\* Only self-supporting persons are included in working population.



5.40. The duration of work in organised industries is governed by the Factories Act, 1948, both in regard to perennial factories which work all the year round and the seasonal factories which work during a particular season like the sugar industry which works during the period sugar-cane is available. The total working days in all industries showed an increase in the year 1950 of the order of 1.8 per cent over the year 1949, the actual figures being 34,34,000 in 1949 as against 34,96,000 in 1950.

5.41. Out of the total working days for all industries, those in the ten major industries formed broadly 20 per cent during both the years. Total man-hours worked for all the industries stood at 519.66 crore and 493.36 crore in 1949 and 1950 respectively. Out of these, the man-hours worked in the ten major industries accounted for 62 and 59 per cent in 1949 and 1950 respectively. The maximum man-hours worked during the two years were in cotton textile, jute and iron and steel industries in the descending order.

5.42. In this survey each manufacturer was asked to report the average number of employees per day receiving pay within the calendar year. The employees were classified into two broad groups, namely, (i) wage earners and (ii) others. 'Employees' include all administrative, technical and clerical staff working within the factory area and all those engaged in effecting delivery of the output. But persons employed in any retail sales organisation maintained by the factory and those engaged in sale of goods which were not subjected to any manufacturing process but merely bought and re-sold have been excluded. The personnel of the central administrative offices outside the factory area have also not been included.

5.43. *Wage earners and wages* : Wage earners in manufacturing plants are, generally speaking, those who perform manual work using tools, operating machines, handling materials and products and care for plant and its equipment. They comprise of both time-workers and piece-workers. Worker does not include a person solely employed in a clerical capacity in any room or place where no manufacturing process is carried on.

5.44. The average number of persons employed per day has been worked out by dividing the aggregate number in attendance on all working days by the total number of working days during the year. In reckoning attendances, attendances by *badli* or substitution and temporary as well as permanent employees have been counted. Total attendances have been arrived at by taking the aggregate of daily attendances in respect of all working days. Absence for a few hours only has not been considered. Total attendances on any day are the total of the attendances in each shift during that day. Days on which the factory was closed and days on which the manufacturing processes were not carried on have not been treated as working days.

5.45. The total amount of wages paid to workers in all industries inclusive of other benefits stood at Rs. 213.40 crore in 1949 and Rs. 199.64 crore in 1950 thus recording a decrease of approximately 6.0 per cent in the latter year. This is shown in Table (5.18).



NATIONAL SAMPLE SURVEY : MANUFACTURING INDUSTRIES, 1949 AND 1950

TABLE (5.18) : ESTIMATES OF WAGES AND SALARIES PAID

item	Rs. (crore)	
	1949	1950
(1)	(2)	(3)
1. wages (inclusive of benefits for workers)	213.40	199.64
2. salaries (inclusive of benefits for persons other than workers)	51.20	53.72

5.46. The total salaries paid to persons other than workers in all industries recorded an increase of 5 per cent in 1950 over the previous year.

5.47. The wages bill inclusive of other benefits in respect of all the workers per working day in the ten major industries was of the order of 65.5 per cent of the total wages including benefits for all industries during the years 1949 and 1950. The highest absolute wage inclusive of other benefits was recorded in cotton textile, followed by that in jute, iron and steel, sugar and tea, the first three alone accounting for about 86 per cent of the wages bill for the ten major industries, and to about 57 per cent of the wages bill for workers for all industries. The rates of remuneration for workers and for persons other than workers are set out in Table (5.19).

TABLE (5.19) : ESTIMATES OF WAGES AND SALARIES FOR WORKERS AND PERSONS OTHER THAN WORKERS IN 1949 AND 1950

industry	workers				other than workers				average earnings	
	wages including benefits		average earning per worker		wages including benefits		average earnings per person		per employed person per working day	
	Rs. (crore)		per day	Rs.	Rs. (crore)		per day	Rs.	Rs.	
	1949	1950	1949	1950	1949	1950	1949	1950	1949	1950
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1. cement	1.59	1.57	2.60	2.76	0.36	0.44	5.18	6.62	2.87	3.17
2. cotton textile	89.18	76.06	4.15	4.26	10.30	10.09	8.22	8.92	4.37	4.52
3. heavy chemicals	3.10	2.99	3.90	3.55	1.40	1.91	8.00	9.57	4.63	4.71
4. iron and steel	8.69	10.29	6.56	7.30	3.34	2.97	11.18	8.65	7.40	7.56
5. jute textile	25.24	22.60	3.20	3.09	3.07	3.13	7.15	7.86	3.40	3.34
6. paints and varnishes	0.23	0.15	2.66	2.26	0.13	0.13	5.76	5.81	3.31	3.17
7. paper and paper board	1.94	1.70	2.87	2.09	0.68	1.22	7.20	10.99	3.40	3.16
8. sugar	6.37	6.09	7.43	5.85	2.31	2.37	10.40	9.41	8.04	6.55
9. tea	4.46	5.71	1.97	2.42	2.19	2.32	6.07	6.18	2.53	2.94
10. tobacco	1.38	1.46	2.47	2.31	0.44	0.54	5.74	6.03	2.87	2.77
11. total (1 to 10)	142.18	128.62	4.37	4.16	24.22	25.12	8.00	8.27	4.68	4.53
12. total of all industries	213.40	199.64	4.45	4.25	51.20	53.72	8.91	9.22	4.92	4.80
13. per cent to total of all industries	66.63	64.42	98.20	97.88	47.30	46.76	89.79	89.70	95.12	94.38



5.48. The highest absolute salaries including benefits for persons other than workers are in cotton textile, jute, iron and steel, sugar and tea. The first three of these account for 69 per cent of the total salaries for the ten major industries and about 33 per cent of the total salaries including benefits for all industries for employees other than workers.

5.49. The average earnings per worker per working day in all industries were Rs. 4.45 in 1949 and Rs. 4.25 in 1950 whereas for persons other than workers the corresponding figures were Rs. 8.91 in 1949 and Rs. 9.22 in 1950. Thus in all industries the average earnings for workers per working day dropped in 1950 by about 5 per cent whereas the earnings increased by 3 per cent for persons other than workers. The overall average earnings per working day for all employed persons recorded, however, a nominal fall in 1950 of the order of 2 per cent.

5.50. In the ten major industries the average wages per worker per working day were Rs. 4.37 in 1949 and Rs. 4.16 in 1950. The average earnings per worker per day in the ten major industries as compared to those in all industries were of the order of 98 per cent, the maximum average earning per worker being in iron and steel, sugar, cotton textiles and jute textile industries in descending order.

5.51. The rate of earnings for workers and for persons other than workers in all industries were Rs. 4.92 in 1949 and Rs. 4.80 in 1950. The comparative average earnings in the ten major industries were Rs. 4.68 in 1949 and Rs. 4.53 in 1950.

5.52. Table (5.20) shows the number of workers in relation to total wages and value of input and output. The total number of workers in all

TABLE (5.20): NUMBER OF WORKERS IN RELATION TO TOTAL WAGES AND VALUE OF INPUT AND OUTPUT IN 1949 AND 1950

industry	number of workers (thousand)		amount received by workers Rs. (crore)		value of input Rs. (crore)		value of output Rs. (crore)	
	1949	1950	1949	1950	1949	1950	1949	1950
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. cement	18.54	16.15	1.59	1.57	6.83	8.47	12.37	16.96
2. chemicals	31.09	31.08	3.10	2.99	11.04	13.15	22.92	27.17
3. cotton textiles	778.46	711.07	89.18	76.06	244.79	258.26	381.33	366.15
4. iron and steel	65.00	63.22	8.69	10.29	29.38	30.55	58.04	57.43
5. jute textile	302.47	283.44	25.24	22.60	109.51	101.58	144.68	148.17
6. paints and varnishes	3.09	2.38	0.23	0.15	2.87	2.45	4.50	3.49
7. paper and paper board	24.46	28.96	1.94	1.70	8.21	9.57	13.69	16.18
8. sugar	100.95	102.05	6.37	6.09	64.36	65.27	95.10	101.17
9. tea	97.14	98.61	4.46	5.71	34.45	37.88	92.20	107.19
10. tobacco	24.08	25.89	1.38	1.46	25.12	28.97	35.32	39.42
11. total (1 to 10)	1445.28	1362.85	142.18	128.62	536.56	556.15	860.15	883.33
12. total of all industries	2424.00	2337.00	213.40	199.64	1072.82	1128.69	1571.39	1642.18
13. per cent to total of all industries	59.62	58.32	66.63	64.42	50.01	49.27	54.74	53.79



# NATIONAL SAMPLE SURVEY : MANUFACTURING INDUSTRIES, 1949 AND 1950

industries comprised 2424 and 2337 thousand during 1949 and 1950 respectively. The amount received by workers in these two years stood for all industries at Rs. 213.40 and Rs. 199.64 crore. The total value of input was Rs. 1072.82 crore in 1949 and Rs. 1128.69 crore in 1950 whereas the respective values of output were Rs. 1571.39 and Rs. 1642.18 crore. Corresponding to that in all industries for the years 1949 and 1950, the number of workers in the ten major industries formed respectively 60 and 58 per cent and the value of output 53 and 54 per cent. Per capita labour earnings in all industries were Rs. 975 in 1949 and Rs. 964 in 1950, the corresponding figures in the ten major industries being Rs. 1053 in 1949 and Rs. 1027 in 1950. Thus in the ten major industries the rate of earnings was low and the average annual earning high compared to all industries because the number of working days in the former case was 226 and 200 in the latter case.

## EMPLOYMENT *VIS-A-VIS* CAPITAL INVESTMENT

5.53. The total invested capital in all industries was Rs. 1068.05 crore in 1949 and Rs. 1100.96 crore in 1950. The corresponding figures of fixed capital are Rs. 538 and Rs. 583 crore respectively. As against this the total number of persons employed in all industries stood at 2714 thousand in 1949 and 2627 thousand in 1950. The figures of output per employed person was Rs. 5790 and Rs. 6251 in 1949 and 1950 respectively. These figures for all industries are shown in Table (5.21).

TABLE (5.21) : NUMBER OF PERSONS EMPLOYED, PER CAPITA COST OF EMPLOYMENT AND GROSS OUTPUT PER EMPLOYED PERSON IN ALL INDUSTRIES IN 1949 AND 1950

item	unit	1949	1950
(1)	(2)	(3)	(4)
1. number of employed persons	(thousand)	2714	2627
2. total invested capital	(Rs. crore)	1068.05	1100.96
3. per capita cost of employment in terms of invested capital	(Rs.)	3935	4191
4. fixed capital	(Rs. crore)	538.00	582.76
5. per capita cost of employment in terms of fixed capital outlay	(Rs.)	1982	2218
6. gross output per employed person	(Rs.)	5790	6251

5.54. The invested capital per employed person in all industries works out to Rs. 3935 in 1949 and Rs. 4191 in 1950. The fixed capital per employed person recorded an increase of 11.9 per cent in 1950 over 1949. The output per employed person stood at Rs. 5790 and Rs. 6251 during 1949 and 1950, which meant an increase of 8 per cent in the latter year.



5.55. An investment in the ten major industries of the order of 55.14 and 54.68 per cent during 1949 and 1950 respectively of the total invested capital in all the industries went towards providing employment broadly to 60 per cent of the total employed persons in all industries. The fixed capital per employed person formed 87 per cent in 1949 and 90 per cent in 1950 of that in all industries. The cost of providing employment to one person in these ten industries varied on an average from Rs. 3700 to Rs. 4000. The corresponding output per employed person lay in the range of Rs. 5400 to Rs. 5900. In the order of capital-intensity per employed person, the ten major industries can be arranged as follows : iron and steel, cement, paper and paper board, tobacco, paints and varnishes, tea, sugar, cotton textile and jute. The highest fixed capital investment among the ten industries was in iron and steel, followed by cement, paper and paper board, chemicals, tea, tobacco, sugar, paints and varnishes, cotton and jute. This is shown in Table (5.22).

TABLE (5.22) : CAPITAL AND OUTPUT PER EMPLOYED PERSON IN 1949 AND 1950

industry	invested capital per employed person Rs.		fixed capital per employed person Rs.		output per employed person Rs.	
	1949	1950	1949	1950	1949	1950
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. cement	8057	8990	5197	5927	5986	9394
2. chemicals	8000	7824	4786	4659	6047	7068
3. cotton textiles	2796	3112	965	1168	4628	4844
4. iron and steel	8521	9042	5483	6073	7286	7304
5. jute textile	2059	2213	838	958	4536	4947
6. paints and varnishes	6920	6754	1845	1720	11532	10913
7. paper and paper board	7701	7343	5409	5447	4910	4919
8. sugar	4579	4519	1625	2106	7481	7982
9. tea	6554	7034	4511	4637	8186	9378
10. tobacco	7976	6508	2337	2313	12899	13352
11. total (1 to 10)	3727	4023	1725	1993	5444	5902
12. total of all industries	3935	4191	1982	2218	5790	6251

5.56. Thus the invested capital in the ten major industries per employed person, recorded an increase of 8 per cent in 1950 over 1949, while the fixed capital was marked by an increment of the order of 16 per cent in the latter year. The output per employed person on the other hand witnessed a rise by about 8.4 per cent in 1950 over the previous year.

#### CAPITAL FORMATION

5.57. The fixed capital in all industries was Rs. 538 crore in 1949 and Rs. 583 crore in 1950. This meant an increase in the fixed capital outlay in all

industries taken together of the order of Rs. 45 crore or 8.36 per cent in course of one year. For the ten major industries on the whole the fixed capital increased in 1950 over 1949 by Rs. 26 crore or by 9.5 per cent. Individually for each of these ten industries there was a general increase in the figures of fixed capital in 1950 over 1949.

#### PRICE FACTOR

5.58. All that has been said above regarding increase in the value of production would mean little real gain as the recorded higher attainments in 1950 may be due to the higher prices in the latter year. The price factor has to be taken into account for final appraisal. It would be well to remember that the index number of wholesale prices was 308 for 1947. After the partial decontrol experiment in that year, the index number of wholesale prices came to be stabilised around 380 in 1948. During the year 1949 this index stood at 385. Thereafter, towards the beginning of 1950, a downward trend in prices was noticeable as a result of a general readjustment of prices all over the world. But this falling tendency received a setback owing to the outbreak of war in Korea in June 1950. Prices went up rapidly and wholesale price index for the year 1950 stood at 409. Prices in 1950 were thus higher to the tune of 6 per cent over those in 1949.

5.59. Assuming that the wholesale index number is applicable to the field under review, we find that for all industries taken together as also for ten major industries there was very little increase in the real value added per worker. Certain individual industries, however, such as cement and jute recorded a real increase of the order of 75 per cent and 35 per cent respectively. Other industries where real increase in the value added per worker was broadly in the neighbourhood of 12 per cent in the year 1950 were chemicals, tea and sugar. On the other hand, the real value added per worker in 1950 as compared to 1949 declined sharply in paints and varnishes followed by cotton textiles, iron and steel, tobacco, and paper and paper boards.

5.60. The rise in prices of the order of 6 per cent in 1950 as compared to the price level in the year 1949 can hardly be taken to usher in anything like a boom in the real sense for industries. It is true that owing to the particular tempo of the Korean war, the products of jute, cement, chemicals and cotton textile industries received a larger assured export market, but the industries found it difficult to expand their production adequately so as to make full use of the favourable market conditions.



## APPENDIX I

INDUSTRY-WISE TABLE SHOWING THE NUMBER OF SAMPLE FACTORIES AND THE TOTAL NUMBER OF FACTORIES COVERED IN SSMI : 1949 AND 1950

industry	total number of sample factories	total number of factories
(1)	(2)	(3)
1. wheat flour	8	115
2. rice milling	38	1574
3. biscuit making	10	70
4. fruit and vegetable processing	6	26
5. sugar : vacuum pan factories	72	171
6. sugar : gur and jaggery refineries	7	7
7. sugar : gur factories	10	253
8. distilleries and breweries	8	66
9. starch	6	22
10. vegetable oils : oil mills	58	1282
11. vegetable oils : hydrogenated	9	34
12. paints and varnishes	8	48
13. soap	11	82
14. tanning	17	115
15. cement	11	17
16. glass and glassware	31	177
17. ceramics	12	76
18. plywood and tea chests	11	47
19. paper and paper board	15	52
20. matches	14	79
21. cotton textiles : spinning mills	38	84
22. cotton textiles : composite mills	232	312
23. cotton textiles : power-loom mills	32	359
24. woollen textiles	16	67
25. jute textiles	100	104
26. chemicals (including drugs etc.)	27	210
27. aluminium, copper and brass : primary products	3	3
28. aluminium and brass etc. : secondary products	26	254
29. iron and steel : primary products	5	5
30. iron and steel : other than primary products	17	198
31. bicycles	6	17
32. sewing machines	6	6
33. producer gas plants	3	3
34. electric lamps	10	10
35. electric fans	9	43

# NATIONAL SAMPLE SURVEY : MANUFACTURING INDUSTRIES, 1949 AND 1950

INDUSTRY-WISE TABLE SHOWING THE NUMBER OF SAMPLE FACTORIES AND THE TOTAL  
NUMBER OF FACTORIES COVERED IN SSME- 1949 AND 1950 (Contd.)

industry	total number of sample factories	total number of factories
(1)	(2)	(3)
36. general and electrical engineering repair workshop	7	210
37. " " " " manufacturing	114	1730
38. footwear and leather manufacturing	8	31
39. rubber and rubber manufacturing	10	99
40. enamelware	6	26
41. hume pipes and other cement and concrete products	7	57
42. asbestos and asbestos cement products	3	5
43. bricks, tiles, lime and surki manufacturing	23	219
44. lac	15	78
45. turpentine and rosin	2	2
46. plastic (including gramophone records)	8	33
47. saw milling	9	305
48. woodware (including furniture)	19	149
49. tea manufacturing	78	1080
50. tobacco products	19	150
51. groundnut decorticating etc.	26	382
52. printing and bookbinding	57	915
53. webbing narrow fabrics	8	86
54. hosiery and other knitted goods	27	240
55. thread and thread ball making	12	36
56. textiles, dyeing, bleaching etc.	19	173
57. clothing and tailoring	11	27
58. cotton ginning and pressing	119	2767
59. rope making	4	8
60. silk and artificial silk	33	313
61. jute pressing	15	39
62. electricity generation and transformation	16	225
63. automobiles and coach building	38	421
64. ship building and ship repairing	14	60
65. aircraft assembling and repair service	7	18
66. railway wagon manufacturing	3	4
67. textile machinery and accessories	17	103
68. unspecified industries	96	1398
69. total of all industries	1742	17377



## APPENDIX II

## FACSIMILE OF THE SCHEDULE OF INVESTIGATION

## SAMPLE SURVEY OF MANUFACTURING INDUSTRIES, 1950

Industry.....	State.....	Region.....	Area.....
Stratum No.....	Serial No. of factory in the Stratum.....		
1. Name and address of the factory.....			
2. Value of Capital employed, as at the close of the year.			
(1) Fixed Capital—			
(11) Land and buildings . . . . .		1949 Rs.	1950 Rs.
(12) Plant and machinery . . . . .			
(13) Other fixed assets . . . . .			
Total . . . . .			
(2) Working Capital—			
(21) Stocks of fuels and raw materials . . . . .			
(22) Stocks of products, by-products, and partly finished products . . . . .			
(23) Cash in hand and at banks . . . . .			
Total . . . . .			
(3) Rent of fixed assets secured on lease . . . . .			
3. Duration of Working—			
(1) No. of days on which any manufacturing operations were carried on . . . . .			
(2) No. of shifts worked per day . . . . .			
(3) Length of Shift . . . . .			

	1949					1950								
	Average No. of persons employed per working day.				No. of manhours worked (workers only).	Amount received by labour.		Average No. of persons employed per working day			No. of manhours worked (workers only)	Amount received by labour		
	Men	Women	Child- ren	Total		Salaries and wages	Value of other benefits	Men	Women	Child- ren		Total	Salaries and wages	Value of other benefits
	Quantity Tons.					Value at factory Rs.		Quantity Tons				Value at factory Rs.		
4. Labour employed —														
(1) Workers	.	.	.	.	.	.	.	.	.	.	.	.	.	
(2) Others	.	.	.	.	.	.	.	.	.	.	.	.	.	
Total	.	.	.	.	.	.	.	.	.	.	.	.	.	
5. Input —														
(1) Fuels, and Electricity—														
(11) Coal and Coke	.	.	.	.	.	.	.	.	.	.	.	.	.	
(12) Other fuels (value only)	.	.	.	.	.	.	.	.	.	.	.	.	.	
(13) Electricity (Electricity generated within the factory should be excluded).	.	.	.	.	.	.	.	.	.	.	.	.	.	
(2) Raw materials, chemicals, packing materials, etc. —														
(21)	.	.	.	.	.	.	.	.	.	.	.	.	.	
(22)	.	.	.	.	.	.	.	.	.	.	.	.	.	
(23)	.	.	.	.	.	.	.	.	.	.	.	.	.	
(24) Others, (value only) including lubricants, consumable stores etc.	.	.	.	.	.	.	.	.	.	.	.	.	.	
(3) Work done for the factory by other concerns (value only)	.	.	.	.	.	.	.	.	.	.	.	.	.	
Total	.	.	.	.	.	.	.	.	.	.	.	.	.	
6. Output—														
(1) Products and by-products—														
(11)	.	.	.	.	.	.	.	.	.	.	.	.	.	
(12)	.	.	.	.	.	.	.	.	.	.	.	.	.	
(13)	.	.	.	.	.	.	.	.	.	.	.	.	.	
(14) Others (value only)	.	.	.	.	.	.	.	.	.	.	.	.	.	
Total	.	.	.	.	.	.	.	.	.	.	.	.	.	
(2) Work done by the factory for customers	.	.	.	.	.	.	.	.	.	.	.	.	.	
Total	.	.	.	.	.	.	.	.	.	.	.	.	.	

Investigator.....

Date.....

Regional Officer.....

Date.....



APPENDIX III

PRINCIPAL PARTICIPANTS

This Survey was initiated in the Directorate of Industrial Statistics under instructions from the Honorary Statistical Adviser to the Cabinet and the Chairman of the National Income Committee and the collection of the data was also made by the staff of that Directorate. Processing and analysis of the data were done and the report (on the basis of a draft prepared by Shri Hari Charan Ghose, then Chief Director of National Sample Survey) prepared by the Indian Statistical Institute.

# THE NATIONAL SAMPLE SURVEY

## NUMBER 12

### A TECHNICAL NOTE ON AGE GROUPING

#### FOREWORD

Biases in age returns occur extensively in India as in many other countries. Attention to this problem has been given in the Indian Censuses; and special groupings have been adopted from time to time for age tabulations and the smoothing of age returns. No systematic study of age distortion has, however, been made so far. It became necessary to consider this question in connexion with the analysis of the demographic data collected in the National Sample Survey (NSS). This technical note gives the results of special investigations undertaken by Ajit Das Gupta and his colleagues in the Indian Statistical Institute for a period of about three years on basis of the NSS and Census age returns, special experiments, and contemporary field studies.

2. The heaping of age returns has been studied in this report for the three components :

- (1) digit preference (as such, without the effects of estimation error and age bias);
- (2) estimation error (as such, without the effects of age bias); and
- (3) age bias;

with a view to isolate the influence of each of these elements by itself (some amount of overlap was, however, unavoidable), and to build up the most efficient set of grouping from the knowledge so obtained.

3. The conventional 0-4 : 5-9 quinary grouping [connoted in the present note by 0 : 5] was found to be relatively inefficient for the NSS data; and the set 2 : 7 came out to be most efficient for important age-income segments of the population : this set also seemed to give a more balanced distribution of the group errors.

4. The superiority of this set was also brought out by other special investigations made by D. B. Lahiri and presented in the paper "Recent developments in the use of techniques for assessment of errors in nation-wide surveys in India" at the International Statistical Conference, Stockholm, 1957. This set had been found to be the most efficient for age returns in the Uttar Pradesh Census of 1951; in the 1931 Census Report also the 2 : 7 grouping had been recommended after an analysis of the age in individual years on traditional lines. No detailed examination could be made for the 1941 Census age data as the tabulations were based on the two per cent Y-sample. In the *Census of India 1951, Paper No. 3, 1954*, some detailed examination of the age data of Uttar Pradesh led to the 2 : 7 set being described as a "standard" grouping; but the 3 : 8 set was recommended as "proper" for reasons not clearly understood.



5. Experience of sample surveys conducted in India suggests that with greater care and interviews at depth, which are not practicable in the Census, it is possible to make improvements in the age returns.

6. The analysis in this report was restricted to the specific objective in view. Other aspects of the quality of population data as obtained through a Census or through the NSS have not been considered here. Studies are, however, going on; and the systematic under-reporting of the population in the young age group 0-14 in the 1951 Census was, for example, examined in *NSS Draft No. 14*, "*Some characteristics of the economically active population*" on the basis of a comparison with age distributions of NSS data on population.

11 October 1958

P. C. MAHALANOBIS

#### ACKNOWLEDGEMENTS

*The Technical Note was prepared by Ajit Das Gupta with the assistance of Samarendra Nath Mitra, and his other colleagues in the Demography Section in the Indian Statistical Institute (ISI).*

*The work in its final form is naturally the product of co-operative labour of men in the Statistical and Field Wings of the National Sample Survey (NSS). Specific mention may be, however, made of Pronoy Kumar Chatterjee for supervision of field work for certain special studies; Jitendra Nath Taluqdar, Gopal Chandra Bhattacharyya and Sukamal Das for supervision of computing; and Suranjn Sen Gupta for editing.*

*Acknowledgement is also due to persons who sent useful comments on the draft.*

# THE NATIONAL SAMPLE SURVEY

## NUMBER 12

### A TECHNICAL NOTE ON AGE GROUPING

#### CONTENTS

	PAGE
FOREWORD ... ..	53
SECTION ONE : Introductory ... ..	57
SECTION TWO : The problem ... ..	59
SECTION THREE : Digit preference ... ..	64
SECTION FOUR : Estimation error ... ..	69
SECTION FIVE : Age bias ... ..	77
SECTION SIX : Measures of concentration and distortion ... ..	80
SECTION SEVEN : Grouping efficiency ... ..	84
APPENDIX 0 : Proforma of Schedule ... ..	88
APPENDIX 1 : Detailed Tables ... ..	89

#### INDEX TO TABLES IN THE TEXT

##### SECTION TWO

TABLE 2.1 : Distribution of individuals by type of available evidence about age .. ..	61
---	----

##### SECTION THREE

TABLE 3.1 : Frequency distribution of the central missing digit supplied by guess.. ..	64
TABLE 3.2 : Frequency distribution of digits supplied by guess in the first two consecutive missing digit places .. ..	65
TABLE 3.3 : Frequency distribution of selected paired consecutive digits supplied by guess .. ..	65
TABLE 3.4 : Frequency distribution of all the three consecutive missing digits supplied by guess .. ..	66
TABLE 3.5 : Population returned at repeated digit individual ages in Census and expected population on elimination of second order of digit preference .. ..	68

##### SECTION FOUR

TABLE 4.1 : Distribution of (1) the second place after decimal of the eye-estimated length of lines and (2) the end-digit of age of all-India rural sample population aged 40-above .. ..	69
---	----



	PAGE
TABLE 4.2 : Distribution of eye-estimate of the aggregate lengths of a cluster of 5 lines (actual aggregate 6.33L) rounded to the first decimal place .. .. .	70
TABLE 4.3 : Distribution of individuals in age-assessed minus age-stated classes under education standard breakdowns .. .. .	70
TABLE 4.4 : Distribution of individuals in age-assessed minus age-stated classes under rating of statement categories .. .. .	71
TABLE 4.5 : Concentration at end-digit '0' in age statements under different rating of statement categories .. .. .	72
TABLE 4.6 : Distribution of individuals in different age ranges under age-assessed minus age-stated groups .. .. .	72
TABLE 4.7 : Distribution of individuals in assessment-evidence type categories under sex breakdowns .. .. .	73
TABLE 4.8 : Concentration at end-digit '0' in age-assessed series under different rating of assessment classes .. .. .	74
TABLE 4.9 : Frequency distribution of the number in different age groups by adjusted difference in ages .. .. .	75

#### SECTION FIVE

TABLE 5.1 : Ratio of numbers returned at each end-digit to total numbers in the successive decennial age ranges .. .. .	78
TABLE 5.2 : First differences of the ratios of numbers returned at each end-digit as shown in Table 5.1 .. .. .	79

#### SECTION SIX

TABLE 6.1 : Measures of concentration at individual end-digits and index of aggregate distortion in age returns .. .. .	81
TABLE 6.2 : Relative range measures of deviation in decennial age ranges .. .. .	83

#### SECTION SEVEN

TABLE 7.1 : Group efficiency index of different sets of grouping .. .. .	84
TABLE 7.2 : Comparative deviations between Census numbers returned and expected under different sets of grouping .. .. .	87

# THE NATIONAL SAMPLE SURVEY

## NUMBER 12

### A TECHNICAL NOTE ON AGE GROUPING

*This Report, A Technical Note on Age Grouping, was prepared by the Indian Statistical Institute and is being published in the form in which it was submitted to the Government of India. The views contained in it are not necessarily those of the Government of India.\**

#### SECTION ONE

#### INTRODUCTORY

1.1. The question of comparative efficiency of different sets of age grouping arose in analysis of National Sample Survey (NSS) demographic data. The feature of heaping up at certain digits, ascribed to 'digit preference', and the resulting distortion of age returns were studied at some depth in this context. The examination of the constituent data itself is no doubt of primary importance in deciding on an efficient set of age grouping, but it was felt that the precise nature of the complex of factors underlying the age distortions had to be understood clearly before the question could be properly tackled. Advantage was, therefore, taken of some experiments contemporarily organised to investigate the interplay of these factors.

1.2. The results of the investigation and the conclusions arrived at are set down in the following sections. The conventional 0-4 : 5-9 grouping, symbolised in the present note as 0 : 5, was found relatively inefficient for the NSS medium and the most efficient set 2 : 7 was adopted in grouping ages for the purpose of internal analysis and also for the purpose of presentation. The departure from the convention itself seemed to call for sufficient justification ; this note was prepared to provide the necessary logical foundation. The findings are of course of wider implication.

1.3. *Summary findings* : The digit preference was examined in isolation from other estimation errors in the Estimation and Extra Sensory Perception (E & ESP) Study 1954 and the examination extended to actual age data. The digit preference as such was seen to have little effect on age record, the estimation errors being by far the dominant factor in the Indian situation where ages were mostly recorded from guess. West Bengal Special Demography (WBSD) Study 1954 for example disclosed that definite evidence of age, including a definite statement of the date of birth and that of children, was available only for one out of six persons, while for about half the population the ages were recorded just from guess.

---

\* The draft report (Number D. 16) was submitted to the Government of India in December 1956.



1.4. The digit preference comprised primarily in a tendency to keep to the middle of the digit array 0, 1, ....., 8, 9 : a liking for a run of consecutive digits and a dislike to repeat digits, for convenience called the second order of digit preference, were also found. The digit preference might be significant in situations where no major distorting factors entered.

1.5. Estimation errors on the other hand produced the familiar pattern of rounding up at digits '0' and '5'. The analysis of the estimation error suggested that apart from the errors of rounding up, there could be a bias to over-estimate. In WBSD Study, both the ages as stated by the informant independently and as assessed by the investigator from the evidence available on his best efforts, were recorded, along with the type of evidence available, the rating of statement and the rating of assessment. The age-assessed was identical with the age-stated in about 3 out of 4 cases; but for the remaining, age-assessed was higher than age-stated nearly twice or thrice as often, more often in the middle age range. The age-assessed series however did not appear to be of any better quality than the age-stated series and over-estimation in assessment was indicated. Due to general ignorance of age, the age assessed by the investigator is usually recorded and the Census age data also supported the finding. The bias to over-estimate the age was confirmed in the West Bengal Household Comparative (WBHC) Study 1955, where the ages of the common population of NSS 4th round and the Study were recorded after a lapse of three years. Significant over-estimation in recorded ages appeared in WBHC Study; the bias actually started as one of under-estimation in the young age range which changed to progressive over-estimation with increasing age, resulting in overestimation in the aggregate.

1.6. The third basic element distorting age returns, the age bias, involving conscious mis-statement of age, was difficult to locate from internal analysis alone, particularly in situations like India where estimation errors are much larger in dimension.

1.7. A modified simple measure of concentration, on the lines of the Myers' index of concentration, was evolved in this note, leading to an index of aggregate distortion; a relative range measure of deviation and a group efficiency index were suggested to enable better analysis and comparative study. It was interesting to note that the average deviation percent of age was nearly uniform in all age ranges, of the order of 0.5. A new technique was applied to determine the most efficient set of age grouping, as the group efficiency index varied for different age segments and socio-economic classes of the same population.



## SECTION TWO

### THE PROBLEM

2.1. While grouping of data is often necessary for presentation and proper comprehension, in the field of age statistics this necessity may be utilised to evolve a set of grouping that reduces the total group errors to a minimum. In a country where ages are as a rule definitely known and reported, the question of the most efficient set of age grouping is not so important, as the group errors will be small for any set. But in a country where ages are generally not definitely known and the heaping up at certain digits at the cost of others is very marked, the selection of an efficient set of grouping is very important.

2.2. This feature of heaping up or concentration at certain popular digits was usually referred as 'integer bias' in the past and sought to be attributed to bias for certain 'preferred' end-digits like 0 and 5. In recent years, however, this is being treated more as an error of rounding off. A good deal of work on the subject of 'integer-bias' or 'round-off' errors in age reporting has already been done, specially in the national census publications of different countries. Age is an important factor not only in the understanding of the vital flows that condition population dynamics but also in sizing up most other population characteristics of socio-economic interest; the need for getting at the best estimate of the true group-age distribution is thus obvious.

2.3. *A priori* considerations suggest that the heaping up in age returns might be the combined effect of the following elements :

- (1) digit preference (as such, without the effects of estimation error and age bias);
- (2) estimation error (as such, without the effects of age bias);
- (3) age bias.

An effort was, therefore, made to grasp the effect of these elements in isolation and to build up the most efficient set of grouping from the knowledge so obtained. The study of these forces in isolation was difficult and some amount of overlapping could not always be avoided.

2.4. From *a priori* considerations again, it would appear that the effect of digit-preference can extend over the unit cycle of end-digits 0, 1, 2, ....., 9 and thus only small displacement errors independent of the age range, should result from it. Estimation error could similarly be expected to produce displacements, small in the earlier age ranges but gradually increasing as age advances. The digit preference and the estimation error again, from their very nature, would be of cyclical nature over the array of end-digits. The age bias, arising as it does from extraneous influences was more likely to have a few focal points at the crucial ages (specific



for different countries in a given period of time), apart from some general tendency to understate or overstate at particular age regions, without any cyclical characteristics : the pull of age bias is apt to be lopsided and to have a long arm.

2.5. If ages are exactly known, stated and recorded the true distribution will of course be reproduced. When ages are exactly known but not correctly stated, age bias will obviously be the element responsible. If ages are known within a narrow margin, the digit preference may conceivably be an important element; but when ages are not known, or only known to lie within widely separated limits, the estimation error is likely to be the dominant element. It is natural that more than one element will be found superimposed on the dominant element in a practical situation, for example when the ages are only known to lie within widely separated limits, the limiting ages themselves will be liable to the influence of age bias.

2.6. What usually happens in a country like India is that the age is unknown and has to be estimated from looks or from comparison with known events or relative seniority ranking within the household or community. Such assessment of age has to be done by the field investigator or enumerator : in reality, an age band with its length depending on the type of evidence available, is consciously or unconsciously estimated by the investigator in the first instance and before the allocation to an individual age within the band. Behind each of the recorded individual ages (falling in the category not definitely known) is, therefore, an estimation age band.

2.7. In WBSD Study<sup>2.1</sup>, among other things, information about the type of evidence available on age was collected, along with the age as stated by the informant, the rating of the statement and the age assessed by the investigator in the field. Information about the type of available evidence is set in Table (2.1).

2.8. It will be seen from Table (2.1) that in West Bengal, where the accuracy of age assessment might be the best for India<sup>2.2</sup>, year of birth of only about 15 per cent of the total population was definitely known. The ages could be estimated or known approximately in about 40 per cent cases; and for the balance of about 45 per cent the age recorded was more or less guess-work estimates. Even in the definite class, documentary evidence of age was obtained in negligible proportion of cases, particularly in the city area. This position should be borne in mind while considering the age returns in the Indian situation.

---

<sup>2.1</sup> This was an experiment on methodology conducted in connection with the NSS. The NSS 4th round sample villages, and the urban and city blocks in West Bengal were adopted for the Study. 744 sample households (hhs.) in 71 villages, 405 sample households in 26 urban blocks and 170 sample households in 14 city blocks (in Calcutta) were interviewed in the Study. Only 18 households had to be substituted, mostly occasioned by subsequent removal. Original NSS sampling fractions were adjusted in a manner as to give uniform multipliers for the three agglomeration sectors. 43 investigators were employed in the experiment and about 1850 investigation-inspection days used up during April-June in the Study.

<sup>2.2</sup> As measured by the Index of Concentration evolved by the U.S. Bureau of the Census and adopted by the Indian Census; *Census of India 1951, Paper No. 3, 1954*, p.4.



# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

TABLE (2.1) : DISTRIBUTION OF INDIVIDUALS BY TYPE OF AVAILABLE EVIDENCE ABOUT AGE

(NSS WBSD Study 1954)

age-assessed group	type of evidence				total
	hoarsay, guess or eye-estimate	related with definite or approximate ages or events	definite statement of year of birth	birth certificate or other documentary	
(1)	(2)	(3)	(4)	(5)	(6)
city (170 households)					
1. 0—6 (%)	26 (27.1)	49 (51.0)	21 (21.9)	—	96 (100.0)
2. 7—16 (%)	40 (41.7)	40 (41.7)	16 (16.6)	—	96 (100.0)
3. 17—above (%)	262 (65.8)	79 (19.8)	54 (13.6)	3 (0.8)	398 (100.0)
4. all ages (%)	328 (55.6)	168 (28.5)	91 (15.4)	3 (0.5)	590 (100.0)
other urban (405 households)					
1. 0—6 (%)	81 (24.6)	138 (42.0)	105 (31.9)	5 (1.5)	329 (100.0)
2. 7—16 (%)	137 (33.9)	199 (49.3)	60 (14.8)	8 (2.0)	404 (100.0)
3. 17—above (%)	581 (53.6)	423 (39.0)	60 (5.5)	20 (1.9)	1084 (100.0)
4. all ages (%)	799 (44.0)	760 (41.8)	225 (12.4)	33 (1.8)	1817 (100.0)
rural (754 households)					
1. 0—6 (%)	129 (18.9)	232 (34.0)	302 (44.2)	20 (2.9)	683 (100.0)
2. 7—16 (%)	322 (36.7)	399 (45.5)	132 (15.1)	24 (2.7)	877 (100.0)
3. 17—above (%)	958 (46.2)	920 (44.4)	137 (6.6)	57 (2.8)	2072 (100.0)
4. all ages (%)	1409 (38.8)	1551 (42.7)	571 (15.7)	101 (2.8)	3632 (100.0)

2.9. Chart (1) gives the histogram representing the distribution of NSS 4th round all-India urban sample population in individual ages, to demonstrate the extent of age distortions involved. The seriousness of the situation will be apparent when it is realised that in the NSS material used in the histogram the population returned at the single individual age 60 is 53.4 per cent of the total population returned in the age group 56–60 and 64.5 per cent of the total returned in age group 60–64; the respective proportions for the rural sector are 61.1 per cent and 68.6 per cent. And the quality of age reporting in NSS medium, as would be anticipated, was somewhat superior to the general census standard. These facts on the type of evidence demonstrate how limited is the possibility of improving the quality of the Indian age returns, perhaps for at least a generation to come.



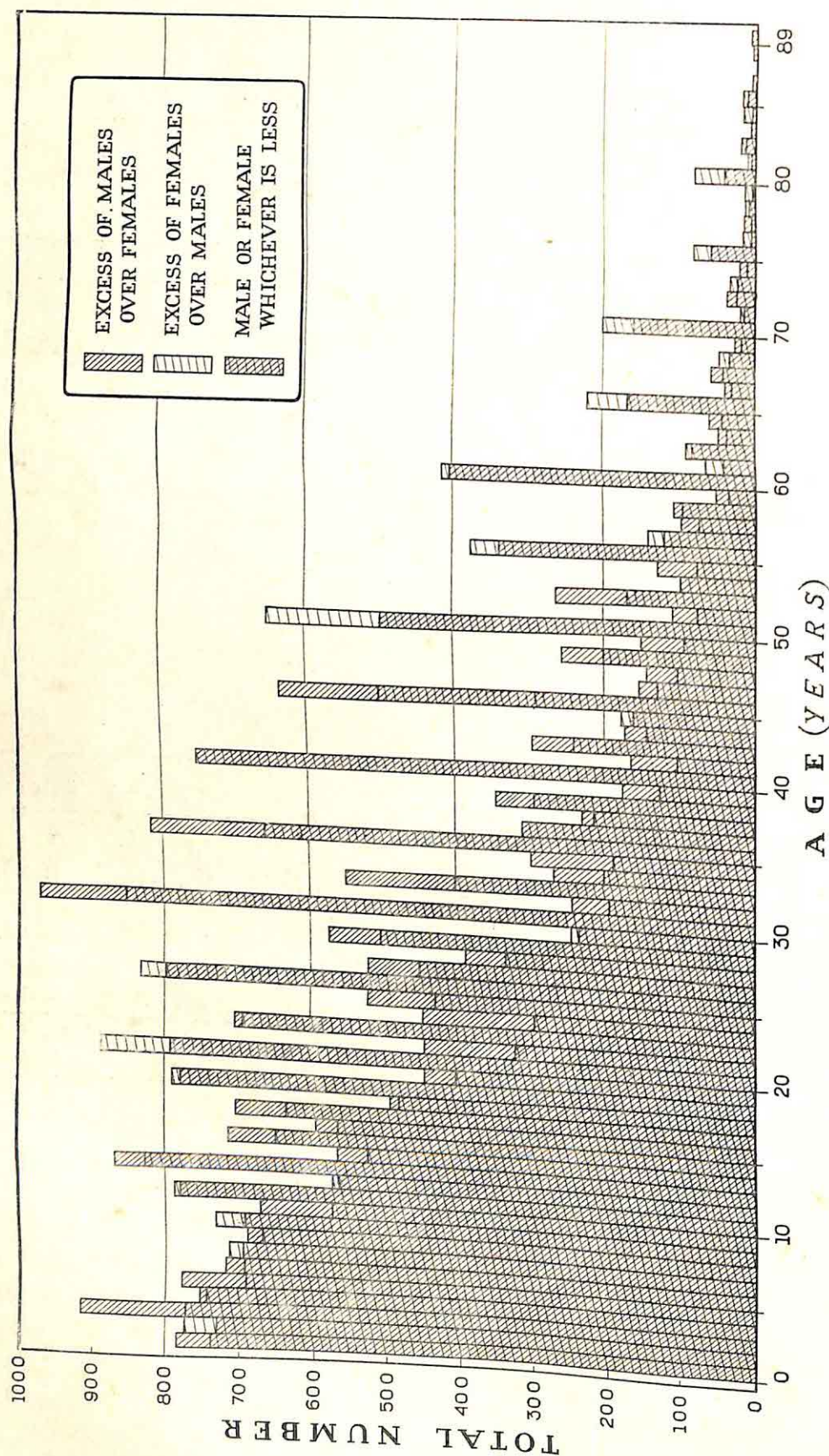


CHART (1) : Frequency distribution of total number returned at each individual age (NSS 4th round, urban).

The number of males or females returned are shown in the diagram on a vertical scale half of that for total numbers. The total population lies exactly half way between the top of the cross-hatch portion and the top of the bar.

## NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

2.10. Discussions will be confined in this note to uniform quinary groups, the smallest groups usually adopted in representation of age statistics. Analysis for a most efficient set of decennial groups will involve similar considerations, though extension of the group interval to cover the full cycle of digits makes the problem somewhat easier here. Systems of unequal groups (alternate ternary—septenary for example) are unusual and inconvenient for presentation purposes. In what follows accordingly by the set of most efficient grouping will be meant the set of uniform quinary groups which gives the best fit to the true conditions over the whole range of life. The chronological age is only dealt with in this note.

2.11. As will be apparent from subsequent discussions a certain set of grouping may not be most efficient both in rural and urban conditions, and for different income slabs within the population. In the same manner, it is conceivable that one single set may not be equally efficient, say, for different economic activity segments like earners and dependants, or for different universe of events like births and deaths. As treatment of different sectors of the population in varying sets of grouping will disturb comparability, compromise is clearly called for.



## SECTION THREE

## DIGIT PREFERENCE

3.1. The nature and extent of the digit preference was studied in isolation in the first instance. In the E & ESP Study 1954 conducted in the Indian Statistical Institute (ISI), a cluster of five 7-digit numbers without the 3 middle digits was given to the workers for supplying the suppressed middle digits by guess. The distribution of the central digit so supplied by 220 workers, expected to be free from any extraneous bias apart from the digit preference as such<sup>3.1</sup>, is given in Table (3.1).

TABLE (3.1): FREQUENCY DISTRIBUTION OF THE CENTRAL MISSING DIGIT SUPPLIED BY GUESS

(ISI, E and ESP Study 1954)

	digit										total
	0	1	2	3	4	5	6	7	8	9	
frequency	86	59	133	117	115	134	127	140	94	95	1100
(%)	(7.8)	(5.4)	(12.1)	(10.6)	(10.4)	(12.2)	(11.6)	(12.7)	(8.6)	(8.6)	(100.0)

with expected frequency 110 in each cell,  $\chi^2 = 54.8$ , significant at 1%.

3.2. It was apparent that true digit preference comprised a tendency to keep to the middle of the digit array 0, 1, ..., 9, within the range 2-7. The shortfall of the actual frequency from the expected (110) was quite marked at the end-digits, 0, 1, 8, 9<sup>3.1</sup>. This pattern of digit preference is altogether different from the 'integer bias' with strong pulls for 0 and 5 and smaller pulls for 2 and 8 that emerges from the analysis of age returns. As will be seen later, the heaping up at 0, 5, 2 and 8 arises mainly from estimation error, which is the most powerful element behind the distortion in age reporting.

3.3. Other interesting facets of digit preference were disclosed by the E & ESP Study<sup>3.1</sup>. One was the disinclination to repeat digits in one sequence and the other the preference for a run of consecutively rising digits; we shall call this the second order of digit preference. Table (3.2) gives the distribution of two consecutive filled-in digits in the E & ESP study (the missing third and fourth places of 7-digit numbers).

<sup>3.1</sup> It is relevant to mention that the effect of extra-sensory perception (ESP) was found negligible. Incidentally, an understanding of digit preference may be usefully employed to check and control (numerical) copying and computational mistakes.



# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

TABLE (3.2): FREQUENCY DISTRIBUTION OF DIGITS SUPPLIED BY GUESS IN THE FIRST TWO CONSECUTIVE MISSING DIGIT PLACES

(ISI, E and ESP Study 1954)

first missing digit	second missing digit										total
	0	1	2	3	4	5	6	7	8	9	
0	9	7	9	4	6	5	1	2	3	3	49
1	14	9	29	13	10	10	7	7	6	3	108
2	12	8	6	22	16	13	18	12	9	7	123
3	6	4	14	6	33	20	15	18	11	14	141
4	11	4	19	24	5	38	32	13	7	8	161
5	9	5	12	14	13	9	24	17	14	6	123
6	5	5	16	14	7	10	7	40	11	7	122
7	5	5	7	5	7	11	13	7	21	13	94
8	5	3	11	6	9	10	4	16	3	21	88
9	10	9	10	9	9	8	6	8	9	13	91
total	86	59	133	117	115	134	127	140	94	95	1100

3.4. Table (3.3) shows separately the frequencies of selected digit pairs of Table (3.2). The expected frequency in each cell of the table on basis of random selection of digits is 11. Some of the lowest frequencies occurred with the repeated digit pairs 00, 11, 22, ....., 99, the total frequency of this set of ten repeated digit numbers being only 74 against expected 110. On the other hand, most of the highest frequencies occurred with the set of eight consecutively rising digit pairs 12, 23, ....., 89, the total frequency of the set being 228 against expected 88.

TABLE (3.3): FREQUENCY DISTRIBUTION OF SELECTED PAIRED CONSECUTIVE DIGITS SUPPLIED BY GUESS

(ISI, E and ESP Study 1954)

consecutive rising run of digits		repeated digits	
selected paired digits	frequency	selected paired digits	frequency
(1)	(2)	(1)	(2)
12	29	00	9
23	22	11	9
34	33	22	6
		33	6
45	38	44	5
56	24	55	9
67	40	66	7
		77	7
78	21	88	3
89	21	99	13
total	228	total	74
average	28.5	average	7.4

with expected frequency 11 in each cell  $\chi^2 = 263.4$  with degrees of freedom 8, significant at 1%.

with expected frequency 11 in each cell  $\chi^2 = 18.2$  with degrees of freedom 10, significant at 5%.



3.5. Similar preferences for run of consecutively rising digits and dislike for the repeated digits were found in the analysis of other filled-in places of the numbers. Part of the short-fall in frequencies of the repeated digit pairs clearly resulted from the attraction for the consecutively rising digit pairs, contiguous to them. The preference is also disclosed in the frequency distribution of all the three filled-in missing digits given in Table (3.4). Arranged in the table in the order in which they occurred in the filled-in E & ESP schedules, the progressive diagonal shift of the maximum frequency range down the table is apparent.

TABLE (3.4) : FREQUENCY DISTRIBUTION OF ALL THE THREE  
CONSECUTIVE MISSING DIGITS SUPPLIED BY  
GUESS

(ISI, E and ESP Study 1954)

digit	third place	fourth place	fifth place
(1)	(2)	(3)	(4)
0	49	86	85
1	108	59	90
2	123	133	88
3	141	117	139
4	161	115	95
5	123	134	143
6	122	127	118
7	94	140	102
8	88	94	116
9	91	95	124

3.6. These inherent likes and dislikes in the run of numbers will naturally enter the reporting by the informant and the assessment and recording by the investigator. The reported individual age distribution of Bengal (males) of Census 1911 and of West Bengal and U.P. (males) of Census 1951 were examined to see how far this second order of digit preference persisted in the census age reporting. The disturbance in age returns from estimation error was much stronger, but the digit preference of the second order being non-cyclic was not masked altogether by this stronger cyclic disturbance. The numbers returned at particular ages under examination could not be compared with the graduated frequency for the purpose of the examination and a method had to be devised to estimate the expected frequency on elimination of the second order of digit preference.

3.7. Chart (2) shows for Bengal (males) 1911 Census population, in the form of smooth curves, the distribution of the numbers returned in individual ages in decennial age-segments.



# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

3.8. It was argued that the form of the frequency curve in the region of an individual age in question, subject to the primary digit preference and estimation error (both of which were of cyclical nature within the digit array) but not subject to the second order of digit preference, will be intermediate to the form of the curve in similar end-digit regions either side, ten years of age up and below. In other

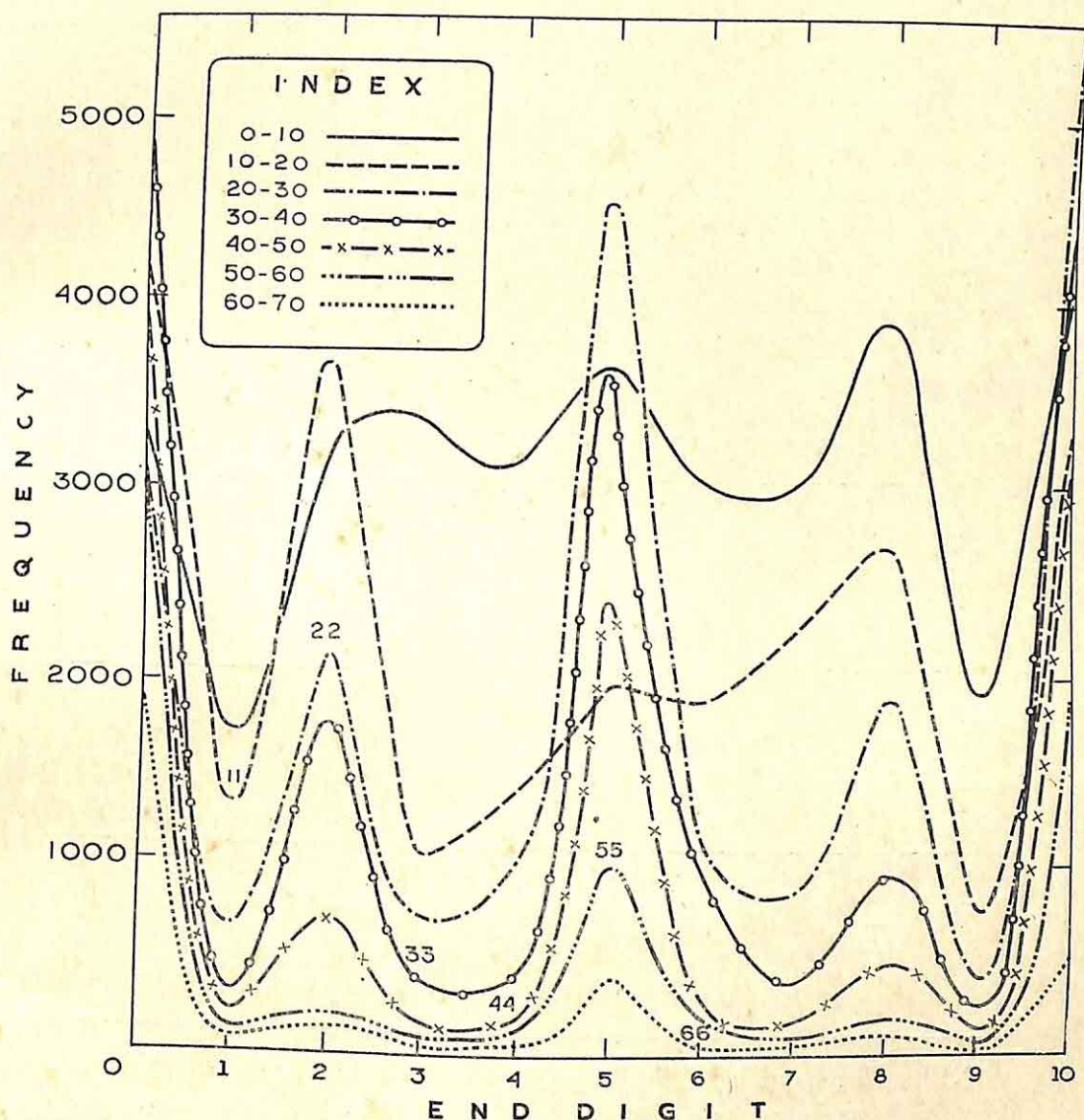


CHART (2) : Frequency curves of numbers returned at each end-digit in decennial age groups (Census 1911, Bengal males).

words, denoting the actual number returned at individual age  $x$ , by the symbol  $n_x$  the form of the individual age frequency curve in the  $n_{10v+u-1} : n_{10v+u} : n_{10v+u+1}$  region would be intermediate between those rendered by the  $n_{10(v-1)+u-1} : n_{10(v-1)+u} : n_{10(v-1)+u+1}$  and  $n_{10(v+1)+u-1} : n_{10(v+1)+u} : n_{10(v+1)+u+1}$  regions, if the second order of digit preference were not there. In effect the assumption allows only for the cyclic distortions and thus eliminates the non-cyclic influences. The distortion from age bias is therefore also not allowed for; but for simplicity the age bias which is localised can be left out of account for the time being.



3.9. The simplest estimations were made in pursuance of the above assumption and constants determined from two linear simultaneous equations on either side were applied to the repeated digit age in question to estimate the appropriate expected frequency. Thus, to estimate  $E(n_r)$  the expected frequency at the repeated digit age  $r = 10v + v$  the formula  $E(n_r) = A_v n_{r-1} + B_v(n_r + n_{r+1})$  was tried, the constants  $A_v, B_v$  being determined from the two equations  $n_{r-10} = A_v n_{r-10-1} + B_v(n_{r-10} + n_{r-10+1})$  and  $n_{r+10} = A_v n_{r+10-1} + B_v(n_{r+10} + n_{r+10+1})$ . Table (2.4) gives the numbers actually returned and the numbers expected at the repeated digit ages, for Census 1911 Bengal (males), and Census 1951 West Bengal (males) and Uttar Pradesh (males).

TABLE (3.5): POPULATION RETURNED AT REPEATED DIGIT INDIVIDUAL AGES IN CENSUS AND EXPECTED POPULATION ON ELIMINATION OF SECOND ORDER OF DIGIT PREFERENCE  
(Census of India)

age $x$		number returned in census (000) $n_x$	expected frequency (000) $E(n_x)$	percentage deviation $\frac{E(n_x) - n_x}{n_x} \times 100$
(1)		(2)	(3)	(4)
Bengal (males) : Census 1911				
1.	11	1310	1606	22.6
2.	22	2156	2304	6.9
3.	33	374	363	-2.9
4.	44	202	218	7.9
5.	55	1017	1096	7.8
6.	66	39	41	5.1
West Bengal (males) : Census 1951				
1.	11	2395	2799	16.9
2.	22	2547	2474	-2.9
3.	33	1397	1491	6.7
4.	44	1250	1292	3.4
5.	55	1068	1119	4.8
6.	66	212	232	9.1
U. P. (males) : Census 1951				
1.	11	6408	8731	36.3
2.	22	6356	6112	-3.8
3.	33	2073	3003	44.9
4.	44	1871	1918	2.5
5.	55	4991	5332	6.8
6.	66	374	406	8.6

3.10. The expected frequency was as a rule higher than numbers returned; the expected frequency was always higher for the repeated digit pairs age 44 above, though small derivations in reverse direction appeared in the younger repeated digit pair ages. As stated earlier, the age bias could not be allowed for in the method of estimation used and the element of age bias perhaps disturbed the expected frequency of the earlier repeated digit pair ages rendered by the method. Use of a broader based formula might have given better balanced estimates. But the evidence in support of the hypothesis that the second order digit preference persists in age reporting was clear enough from the analysis done above.

3.11. It seems probable that the inflation noticed at age 60 in the age returns of many countries may be, at least partly, due to the dislike of the repeated digit number 55.

## SECTION FOUR

## ESTIMATION ERROR

4.1. Some results obtained in the E & ESP Study relevant to the estimation error are discussed first. The E & ESP Study schedule also contained a cluster of five lines, of which the lengths were required to be eye-estimated and recorded to the second place of decimal in terms of an unconventional unit of length 'L' shown on the body of the schedule<sup>4.1</sup>. The conditions were such that the second decimal figure could be nothing better than pure guess. The distribution of the second decimal figure supplied by 222 workers is given in Table (4.1).

TABLE (4.1): DISTRIBUTION OF (1) THE SECOND PLACE AFTER DECIMAL OF THE EYE-ESTIMATED LENGTH OF LINES AND (2) THE END-DIGIT OF AGE OF ALL-INDIA RURAL SAMPLE POPULATION AGED 40-ABOVE

(ISI, E and ESP Study 1954 and NSS 4th round<sup>4.2</sup> 1952)

item	digit										total
	0	1	2	3	4	5	6	7	8	9	
1. second decimal place of estimate (%)	414 (37.3)	25 (2.3)	72 (6.5)	45 (4.0)	38 (3.4)	344 (31.0)	58 (5.2)	41 (3.7)	48 (4.3)	25 (2.3)	1110 (100.0)
2. end digit, age 40-above	2326	276	552	279	297	1265	301	234	352	165	6047
(concentration) <sup>4.3</sup>	(31.3)	(4.3)	(8.8)	(4.7)	(5.2)	(22.7)	(6.2)	(5.1)	(7.8)	(3.9)	(100.0)

4.2. The concentration at 'preferred' digits is now of the familiar pattern found in age returns, though more accentuated here for the digits '5' and '0'. The striking similarity between the frequency distribution of the digit in the second decimal place in the estimated lengths of lines in the E & ESP Study and the end digit of age of the all-India rural sample population aged 40 and above, suggested that most of the error in age returns is that of estimation when the unit digit of age was just a matter of guess.

4.3. On further analysis, a tendency to over-estimate was also disclosed by the E & ESP Study which was suggestive. The actual aggregate length of the five lines correct to the first place of decimals was 6.3L; but the mean of the estimates recorded by the workers was 6.7L. The distribution of the recorded estimates is given in Table (4.2). The over-estimation was highly significant; as against only 27% who came on the side of under to correct estimation, 73% over-estimated the aggregate length. The major peak of the distribution of estimates was at 6.6L.

<sup>4.1</sup> Appendix 0.

<sup>4.2</sup> All the NSS 4th round rural tables of this note cover the six 1/16th part samples 3, 4, 7, 8, 15 & 16 split at the village level for operational convenience.

<sup>4.3</sup> The measure of concentration is a percentage distribution of the end digits defined in para 6.5.



TABLE (4.2): DISTRIBUTION OF EYE-ESTIMATE OF THE AGGREGATE LENGTHS OF A CLUSTER OF 5 LINES (ACTUAL AGGREGATE 6.33L) ROUNDED TO THE FIRST DECIMAL PLACE

(ISI, E and ESP Study 1954)

length (L) frequency			length (L) frequency			length (L) frequency		
(1)		(2)	(1)		(2)	(1)		(2)
1.	5.1	1	11.	6.4	16	22.	7.5	2
2.	5.4	1	12.	6.5	20	23.	7.6	4
3.	5.7	1	13.	6.6	29	24.	7.7	1
4.	5.8	1	14.	6.7	11	25.	7.8	2
5.	6.0	11	15.	6.8	15	26.	7.9	3
6.	6.1	6	16.	6.9	14	27.	8.0	2
7.	6.2	15	17.	7.0	16	28.	8.1	1
8.	6.3	24	18.	7.1	7	29.	8.2	1
9. sub-total: 6.3 below 60			19.	7.2	7	30.	8.3	1
(%) (27.0)			20.	7.3	3	31.	8.4	1
10. sub-total: 6.4 above 162			21.	7.4	6	32.	total	222
(%) (73.0)							(%)	(100)
mean = 6.69			$\sigma^2 = .2627$			$s_m = .0434$		
						$t = 8.3$		

4.4. Such tendency of over-estimation (or under-estimation) has been found by other operators in different experiments<sup>4.4</sup>.

4.5. It was therefore decided to investigate how far the bias to over-estimate or under-estimate might have entered the assessment of age. Actual age data of the WBSD Study were analysed to investigate this. Table (4.3) gives the distribution of the individuals covered by the Study in age-assessed minus age-stated groups, separately for the city, other urban and rural sectors<sup>4.5</sup>.

TABLE (4.3): DISTRIBUTION OF INDIVIDUALS IN AGE-ASSESSED MINUS AGE-STATED CLASSES, UNDER EDUCATION STANDARD BREAKDOWNS

(NSS, WBSD Study 1954)

age-assessed minus age-stated	city (170 households)				other urban (405 households)				rural (754 households)			
	total	illi- terate	liter- ate	matri- culate	total	illi- terate	liter- ate	matri- culate	total	illi- terate	liter- ate	matri- culate
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1. age-assessed = age-stated (%)	495 (83.9)	194 (86.2)	236 (81.9)	65 (84.4)	1552 (86.2)	749 (81.6)	710 (90.5)	93 (94.9)	2733 (77.6)	2012 (76.3)	701 (81.1)	20 (95.2)
2. age-assessed < age-stated (%)	20 (3.4)	8 (3.6)	10 (3.5)	2 (2.6)	70 (3.9)	59 (6.4)	11 (1.4)	— (—)	274 (7.8)	218 (8.3)	56 (6.5)	— (—)
3. age-assessed > age-stated (%)	75 (12.7)	23 (10.2)	42 (14.6)	10 (13.0)	178 (9.9)	110 (12.0)	63 (8.1)	5 (5.1)	515 (14.6)	407 (15.4)	107 (12.4)	1 (4.8)
4. total (%)	590 (100)	225 (100)	288 (100)	77 (100)	1800 (100)	918 (100)	784 (100)	98 (100)	3522 (100)	2637 (100)	864 (100)	21 (100)

<sup>4.4</sup> Examples of such bias of over-estimation in selection of 'representative' units have been cited by Frank Yates in "Sampling Methods for Censuses and Surveys" (1953) at pp. 12-13.

<sup>4.5</sup> Age was not stated in less than 1% of the cases only (most of it in the rural sector) and the not-stated cases have been left out of the tables.



# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

4.6. The age-assessed minus age-stated was positive two to four times more often than it was negative : that is, a higher age was assessed two to four times more frequently. With regard to this feature the male female differential was not significant, but the relative proportion of higher age-assessed tended to go up among non-Hindus and among Hindi-speaking population in the West Bengal field<sup>4.6</sup>. The proportion, was, if anything, rather higher in city area and among the educated, as Table (4.3) shows.

4.7. The investigator's assessment of age is usually the only thing available and recorded in census and surveys in India but in the WBSD Study the investigators were definitely instructed not to render any such assistance in age statement and clear evidence is thus furnished that the investigator assesses a higher age in the sum than the informant states. The question is whether the investigator was trying to correct in his assessment a real under-statement of age by the informant, or if in the aggregate there was no such under-statement in the process of assessment, the age was over-estimated by the investigator. An attempt was made to answer this question from further examination of the WBSD Study material. Table (4.4) gives the distribution of individuals in age-assessed minus age-stated classes for different categories of rating of statement.

TABLE (4.4): DISTRIBUTION OF INDIVIDUALS IN AGE-ASSESSED MINUS AGE-STATED CLASSES UNDER RATING OF STATEMENT CATEGORIES  
(NSS, WBSD Study 1954)

age-assessed minus age-stated	rating of age statement								
	city (170 households)			other urban (405 households)			rural (754 households)		
	guess	appro- ximate	definite	guess	appro- ximate	definite	guess	appro- ximate	definite
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. age-assessed =age-stated (%)	180 (79.3)	153 (81.8)	162 (92.0)	350 (74.2)	513 (82.7)	689 (97.3)	733 (63.3)	992 (75.4)	1008 (96.3)
2. age-assessed < age-stated (%)	11 (4.8)	5 (2.7)	4 (2.3)	38 (8.0)	28 (4.5)	4 (0.6)	142 (12.2)	120 (9.1)	12 (1.1)
3. age-assessed > age-stated (%)	36 (15.9)	29 (15.5)	10 (5.7)	84 (17.8)	79 (12.8)	15 (2.1)	284 (24.5)	204 (15.5)	27 (2.6)
4. total (%)	227 (100.0)	187 (100.0)	176 (100.0)	472 (100.0)	620 (100.0)	708 (100.0)	1159 (100.0)	1316 (100.0)	1047 (100.0)

4.8. For all categories of rating, the proportion of ages assessed higher to ages assessed lower is fairly stable, about 2 for all sectors combined, rather a little higher for the category of rating definite. A progressive fall in the proportion was to be expected in passing from guess to definite category of rating of age statement if the investigator, in his assessment, was correcting under-statements for which the guess and approximate categories offered a much bigger scope. The actual pattern brought out suggests over-estimation in age-assessed.

4.9. The assumption is, however, implicit here that the rating of statement has been proper. As to the reliability of the rating, Table (4.5) gives the frequency



of end-digit '0' in the age-stated range 23-62; a systematic fall in the concentration at end-digit '0' with upgrading in rating category is observed. It is permissible to take this as a very rough test of the validity of rating done.

TABLE (4.5): CONCENTRATION AT END-DIGIT '0' IN AGE STATEMENTS UNDER DIFFERENT RATING OF STATEMENT CATEGORIES

(NSS, WBSD Study 1954)

rating of statement	city (170 households)			other urban (405 households)			rural (754 households)		
	population aged 23-62		concen- tration	population aged 23-62		concen- tration	population aged 23-62		concen- tration
	return- ed at end-digit '0' ages	total	$\frac{\text{col(2)}}{\text{col(3)}} \times 100$	return- ed at end-digit '0' ages	total	$\frac{\text{col(5)}}{\text{col(6)}} \times 100$	retrun- ed at end-digit '0' ages	total	$\frac{\text{col(8)}}{\text{col(9)}} \times 100$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. guess	37	161	23.0	71	226	31.4	160	596	26.9
2. approximate	14	84	16.6	70	341	20.5	143	665	21.5
3. definite	5	51	9.8	36	197	18.3	39	219	17.8
4. total	56	296	18.9	177	764	23.2	342	1480	23.1

4.10. Table (4.6) gives distribution of the gap between age-assessed and age-stated in broad age-assessed groups.

TABLE (4.6): DISTRIBUTION OF INDIVIDUALS IN DIFFERENT AGE RANGES UNDER AGE-ASSESSED MINUS AGE-STATE GROUPS

(NSS, WBSD Study 1954)

age-assessed minus age-stated	age-assessed (years)								
	city (170 households)			other urban (405 households)			rural (754 households)		
	0-16	17-61	62- above	0-16	17-61	62- above	0-16	17-61	62- above
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. -11 below	—	—	1	—	2	—	—	9	2
2. -10 to -6	—	1	—	—	3	—	—	28	2
3. -5 to -3	—	2	2	5	14	3	13	58	8
4. -2 to -1	5	8	1	20	23	—	73	76	4
5. -1 & below (%)	5 (2.6)	11 (2.9)	4 (16.0)	25 (3.4)	42 (4.2)	3 (4.8)	87 (5.7)	171 (9.1)	16 (14.7)
6. 0 (%)	182 (94.8)	296 (79.4)	17 (68.0)	684 (93.3)	815 (81.1)	53 (85.5)	1363 (88.6)	1306 (69.7)	64 (58.7)
7. 1 to 2	5	42	2	20	82	—	75	159	6
8. 3 to 5	—	21	2	4	54	6	13	183	13
9. 6 to 10	—	3	—	—	8	—	—	49	8
10. 11 above	—	—	—	—	4	—	—	7	2
11. 1 & above (%)	5 (2.6)	66 (17.7)	4 (16.0)	24 (3.3)	148 (14.7)	6 (9.7)	88 (5.7)	398 (21.2)	29 (26.6)
12. total (%)	192 (100.0)	373 (100.0)	25 (100.0)	733 (100.0)	1005 (100.0)	62 (100.0)	1538 (100.0)	1875 (100.0)	109 (100.0)



# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

4.11. While a tendency to overstate age in the old age range is well known and is likely to have brought down the proportion of the total over-assessment to the total under-assessment in the old age range 62-above, the even break of over-assessment and under-assessment in the young age range 0-16 is not so easily explained: the margin available for under-statement is however limited in the young age range by the age attained. The spread up of the gap between age-assessed and age-stated is interesting; less than half of the deviations exceed 2 years of age and most of it fall within the limits of  $\pm 5$  years.

4.12. In the WBSD Study, as already indicated, information was collected about the type of evidence available to the investigator in assessment of ages, on his best efforts. Table (4.7) is an alternative presentation of the information obtained in this respect; both the rating of assessment and type of evidence on which the assessment naturally rested were combined here to give composite categories for assessment evidence types. Table (4.7) shows that definite evidence of age was available in 18-30% of cases only; and, as was seen from Table (2.1), a definite statement of age by the informant was behind most of it. Considering that 16-19% of the individuals covered were in the age range 0-6, the grave weakness in the field of the age assessment is quite apparent. The proportions with definite assessment-evidence type were higher than the respective proportions with definite evidence of age available given in Table (2.1), and the gap increased progressively in passing from the city to the rural sector. The age-assessed series could not be taken as a better approximation to the true ages, in the circumstances.

TABLE (4.7): DISTRIBUTION OF INDIVIDUALS IN ASSESSMENT-EVIDENCE TYPE CATEGORIES UNDER SEX BREAKDOWNS

(NSS, WBSD Study 1954)

assessment-evidence type	city (170 households)			other urban (405 households)			rural (754 households)		
	total	male	female	total	male	female	total	male	female
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. guess (%)	276 (46.8)	154 (45.2)	122 (49.0)	372 (20.5)	204 (21.0)	168 (19.8)	883 (24.3)	434 (23.2)	449 (25.5)
2. approximate (%)	191 (32.4)	108 (31.7)	83 (33.3)	1025 (56.4)	521 (53.8)	504 (59.4)	1743 (48.0)	876 (46.7)	867 (49.3)
3. definite (%)	123 (20.8)	79 (23.1)	44 (17.7)	420 (23.1)	244 (25.2)	176 (20.8)	1006 (27.7)	564 (30.1)	442 (25.2)
4. total (%)	590 (100.0)	341 (100.0)	249 (100.0)	1817 (100.0)	969 (100.0)	848 (100.0)	3632 (100.0)	1874 (100.0)	1758 (100.0)



4.13. The values of concentration at end digit '0' under different ratings of assessment, similar to those shown in Table (4.5) but for the age-assessed series now, are given in Table (4.8). Comparative study of the concentration makes it clear that the quality of the age-assessed series is no better than the age-stated series.

TABLE (4.8): CONCENTRATION AT END-DIGIT '0' IN AGE-ASSESSED SERIES UNDER DIFFERENT RATING OF ASSESSMENT CLASSES

(NSS, WBSD Study 1954)

1951

rating of assessment	city (170 households)			other urban (405 households)			rural (754 households)			
	population aged 23-62		concentra- tion	population aged 23-62		concentra- tion	population aged 23-62		concentra- tion	
	return- ed at end digit '0' ages	total	$\frac{\text{col(2)}}{\text{col(3)}} \times 100$	return- ed at end digit '0' ages	total	$\frac{\text{col(5)}}{\text{col(6)}} \times 100$	return- ed at end digit '0' ages	total	$\frac{\text{col(8)}}{\text{col(9)}} \times 100$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. guess	39	150	26.0	49	207	23.7	128	442	29.0	
2. approximate	14	101	13.9	97	437	22.2	204	889	23.0	
3. definite	3	46	6.5	26	140	18.6	44	273	16.1	
4. total	56	297	18.9	172	784	21.9	376	1604	23.5	

4.14. The probable reason why '0' and '5' happen to be the most favoured digits, in that order, can be examined here. Rounding off at '0' naturally gets first preference as one digit, the unit place, is cut out by such approximation; and at that level of approximation, the mid-way digit '5' gets the next natural preference when the estimate is far off from the rounded up ages at the tenth place on either side. The digit preference as such may also be partially responsible for the popularity for the middle of the array digit '5'. After '0' and '5' there is usually a slight preference for even digit over odd: this also has a simple explanation. If an estimate is above '0' but not far away removed from '0', it will be transferred and recorded under '0' and not under '1'; only when it is far away removed from '0' it will be recorded differently, and will then rather be jumped to '2'. The preference for digit '8' has exactly a similar explanation.

4.15. "Census of India 1951—Age Tables"<sup>4.7</sup> gives a diagonal Table for the Madras (male) population showing estimated percentage 'under-statements' and 'over-statements' at each age from 6 to 67. This Table envisages notional transfers only to adjoining ages, and the graduated individual age frequencies taken as true are derived from the set of grouping centred round the most preferred digits '0' and '5'. The set of grouping does not take into account the tendency of over-estimation. In the age range 27-67 of this table, the average over-statement of age is 29 per cent

<sup>4.7</sup> Census of India, 1951, Paper No. 3, 1954, pp. 16-18.



# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

as against the average under-statement of about 20 per cent. In the age range 6-67, the proportions of similar average over-statement and under-statement are 22 per cent and 19 per cent respectively. Supplementary evidence of the tendency of over-estimation in age return is thus disclosed by the Census reporting itself.

4.16. The analysis done above shows that the estimation error really falls into two parts, one arising from rounding off approximation and other from over-estimation. The element of over-estimation missed attention in the past and the estimation error was equated to the error of rounding off.

4.17. In NSS experimental West Bengal Household Comparative (WBHC) Study 1955 the same sample households as of NSS 4th round were revisited after a lapse of about 3 years to measure changes in living conditions during the intervening period : this opportunity was utilised to investigate further the over-estimation bias in age reporting and the ages of the household members were independently ascertained again. Comparisons between ages reported in NSS 4th round and WBHC Study showed some interesting features. Table (4.9) gives the distribution of the deviation between the age reported in the WBHC Study and the age expected on the basis of the three-year old NSS 4th round age return.

TABLE (4.9) : FREQUENCY DISTRIBUTION OF THE NUMBER IN DIFFERENT AGE GROUPS BY ADJUSTED DIFFERENCE IN AGES

(NSS 4th round 1952 and WBHC Study 1955 : 650 rural households)

age difference : WBHCS - (4th round + 3)	age (years) WBHCS					
	3-9	10-19	20-29	30-39	40-above	all ages
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. 0 (%)	184 (45.8)	158 (36.9)	128 (31.8)	84 (27.5)	116 (23.0)	670 (32.8)
2. 1 & above (%)	91 (22.6)	126 (29.4)	134 (33.2)	125 (41.0)	244 (48.4)	720 (35.3)
3. mean difference	1.28	1.81	2.32	2.46	3.17	2.41
4. -1 & below (%)	127 (31.6)	144 (33.7)	141 (35.0)	96 (31.5)	144 (28.6)	652 (31.9)
5. mean difference	-1.67	-1.61	-1.95	-2.54	-2.79	-2.09
6. total (%)	402 (100.0)	428 (100.0)	403 (100.0)	305 (100.0)	504 (100.0)	2,042 (100.0)
7. mean difference	-0.24	-0.01	0.09	0.21	0.74	0.18

with mean difference 0.18,  $t = 3.23$ , significant at 1%.

4.18. The ages reported in WBHC Study were in sum somewhat higher than the ages expected on basis of the three-year old NSS 4th round returns, with an average over-statement of 0.18 years. But this was for the persons common in the two surveys : those born since NSS 4th round survey and those dead since, were naturally excluded from the comparison, apart from the migrants. The average report-



ed ages of the two surveys did not differ significantly. It was interesting to observe that while there was a big comparative under-statement in the youngest present age range 3-9, the difference narrowed in the age range 10-19, then changed to slight overstatement in the next higher age range 20-29 and to increasing over-statements in the later age ranges 30-39 and 40-above. The over-statement observed for the total range was thus the contributory effect of high over-statements at the advanced ages.

4.19. The investigating staff employed in both the surveys were by and large similar in instruction, training and experience; there were therefore no reasons to anticipate any material investigator differences. The reporting population was also more or less the same: they were not subject to repeated surveys in the intervening period nor otherwise conditioned to change. The WBHC Study results thus suggest that there is some under-estimation of advance of ages in the young age ranges below 20 and distinct over-estimation of the advance at the later adult ages, with a moderate over-estimation of ages in the balance. Thus in the aggregate the age of the population is over-estimated: the aggregate over-estimate may remain stable if the old, whose ages were over-estimated most in the previous survey, die in such proportion as to offset the subsequent increased over-estimation of those who live to grow older.



# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

## SECTION FIVE

### AGE BIAS

5.1. The possible location and nature of the age bias in different population formations can often be known from their cultural traits. Tendencies to exaggerate ages in the threshold of adulthood and after retirement, conscious understatement of age in the young-adult range by the females in some culture and conscious over-statement of age to attain legal majority, to escape military service and to qualify for old-age pensions or just to impress as outstandingly old, are some of the common sources of the bias experienced. The mores and the laws of the land work behind the bias, and changes in them deflect the pattern of the bias : the pattern is usually quite stable over time in each country until the relevant laws change. The age bias is of specific location and is thus distinguished from the general estimation bias, from which it also differs in nature.

5.2. Longitudinal comparisons across consecutive census intervals, allowing for migration if significant, and for the digit preference and the estimation error, might disclose the pattern of age bias : except for the digit preference of the second order, these preferences and errors of cyclical nature get automatically allowed for to a large extent in comparisons over a series of decennial censuses. If reliable birth-death registration and migration statistics be available, the total distortions could be determined by reconciliation of successive census results with the intervening movements and the extent of age bias broadly assessed. Such techniques are however not helpful when the relevant data are grossly defective, as in India. Sample verification of ages in the field with a superior set of investigators, who travel down to the birth certificate or other best available evidence of age, is another alternative method employed to locate age bias or rather the total distortions in age recording. It may be reiterated that the cardinal point of interest in the problem of age grouping is the total distortion, and the elements leading to it are studied to get a better understanding of the position.

5.3. It should have been possible to spot the age bias at analysis stage by examining the run of the ratios that the numbers returned at each end digit of age constitute of the total returned in the successive decennial age ranges say. But such examination is also not very helpful for the Indian situation where the big distortions from estimation error mask most other features. Table (5.1) shows for the NSS medium the ratios

$$\frac{n_{10v+u}}{\sum_{u=0}^9 n_{10v+u}} \times 100$$

up to age 79, for each end digit  $u$  in the whole sequence of age ranges

$$\sum_{u=0}^9 n_{10v+u}, \quad v = 0, 1, \dots, 7.$$



TABLE (5.1) : RATIO OF NUMBERS RETURNED AT EACH END-DIGIT TO TOTAL NUMBERS IN THE SUCCESSIVE DECENNIAL AGE RANGES

(NSS 4th round 1952, All-India rural sample : 28,918 persons)

end digit	decennial age range							
	0—9	10—19	20—29	30—39	40—49	50—59	60—69	70—79
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	10.6	15.2	18.7	26.6	31.4	38.6	49.2	51.0
1	10.1	8.1	6.1	4.6	4.5	5.0	4.6	3.8
2	10.9	14.4	13.5	11.8	9.5	9.2	8.9	7.7
3	10.8	8.7	6.9	5.8	5.0	4.5	3.8	4.4
4	9.7	9.4	9.1	5.8	5.6	4.6	4.7	3.0
5	10.9	10.4	18.4	20.4	23.1	19.6	18.7	19.7
6	10.3	11.3	8.4	8.3	6.2	5.0	2.7	3.3
7	8.9	5.8	5.4	4.5	4.9	3.8	2.1	2.7
8	10.2	11.7	9.8	8.4	6.8	6.5	3.6	3.0
9	7.6	5.0	3.7	3.8	3.0	3.2	1.7	1.4
total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

5.4. The ratios for any one end digit of age could have been expected ordinarily to form a smooth progression over the successive decennial age ranges, on which the impact of the age bias (ignoring the comparatively small influence of the second order of the digit preferences) will have produced marked local disturbances. But in any case, the history of the particular population growth and the local mores and laws to be turned to for confirmation : past fluctuations in birth-death experience may also produce isolated tracts of accumulation, particularly in small populations. Sharp rises in the ratios at ages 25 and 60 are observable in Table (5.1). The pull for age 60 was serious enough in the NSS medium to take the quinary age group 60-64 total beyond the 55-59 total. The examination of the run of the ratios as a method of spotting possible age bias is not satisfactory when the concentration at particular end digits is so high and the relative concentrations between the end digits change so violently as in Table (5.1).

5.5. The moot question in the present case was whether these sharp rises in the run of ratios came rather from mounting concentration at these preferred end digits. The age bias operated in a manner so as to accelerate and decelerate the flow of numbers returned in the immediate neighbourhoods of certain crucial ages : yet another approach of spotting the location of age bias suggested itself from this. Examination of the first differences of the ratios over a few consecutive end digits in neighbouring decennial age ranges should show up the acceleration and deceleration effects. Table (5.2) gives the first differences of the ratios in Table (5.1).

# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

TABLE (5.2) : FIRST DIFFERENCES OF THE RATIOS OF NUMBERS RETURNED AT EACH END-DIGIT AS SHOWN IN TABLE (5.1)

(NSS 4th round 1952, all-India rural sample)

end digit	decennial age range							
	0—9	10—19	20—29	30—39	40—49	50—59	60—69	70—79
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
0	4.6	3.5	7.9	4.8	7.2	10.6	1.8	
1	-2.0	-2.0	-1.5	-0.1	0.5	-0.4	-0.8	
2	3.5	-0.9	-1.7	-2.3	-0.3	-0.3	-1.2	
3	-2.1	1.8	-1.1	-0.8	-0.5	-0.7	0.6	
4	-0.3	-0.3	-3.3	-0.2	-1.0	0.1	-1.7	
5	-0.5	8.0	2.0	2.7	-3.5	-0.9	1.0	
6	1.0	-2.9	-0.1	-2.1	-1.2	-2.3	0.6	
7	-3.1	-0.4	-0.9	0.4	-1.1	-1.7	0.6	
8	1.5	-1.9	-1.4	-1.6	-0.3	-2.9	-0.6	
9	-2.6	-1.3	0.1	-0.8	0.2	-1.5	-0.3	

5.6. The differences for a number of consecutive end digits immediately before age 60 are all negative and comparatively larger, and the differences generally change sign in this area. The conditions in the immediate neighbourhood of age 25 are even less distinctive. Some age bias is however now suggested for age 16, which appears to have gained at the cost of ages 13-15. But the evidence is far from conclusive; the big and mounting concentrations at the preferred digits are no doubt mainly responsible for this lack of conclusiveness. And specific localised age bias is also relatively milder in India.



## SECTION SIX

## MEASURES OF CONCENTRATION AND DISTORTION

6.1. It is obvious that the extent of concentration at each digit and the resulting aggregate distortion have to be measured before any efficient set of grouping could be built up. The concentration at each digit can be measured by comparison of deviations of actual numbers returned at individual ages from the expected numbers in the corresponding ages obtained by graduation. It will be assumed however that no graduation of the age returns has been done, as the problem of grouping ceases to exist if a good unbiased graduation, influenced by subjective choice of the operator, is already available; any set of grouping will be equally efficient when built up from such graduated numbers.

6.2. At one time, the total numbers returned at each end digit used to be compared to a tenth of the population, to get a measure of the integer bias and of the total distortion, on the assumption that all the end digits should occur with equal frequency if there was no integer bias. King (1916) then pointed out that the starting integers 0, 1, 2, ..... got additional weightage in that order in such comparison<sup>6.1</sup>. The difficulty was resolved by Myers (1940) who used a blended population for the purpose in his index of concentration<sup>6.2</sup>; each digit was put successively at each place from first to tenth in the component populations by Myers, so that they got balanced weight in the resulting blended population. Myers started with age 10 and showed that the average concentration values for the United Kingdom 1911 Census age returns yielded by his method were remarkably close to the values obtained by King by comparison of the numbers returned against the graduated numbers.

6.3. A much simpler measure of concentration was evolved in connection with the analysis of distortion in NSS age returns. In a normal population structure, where the numbers alive gradually fall with age, digit '0' will show up a higher concentration than really attached to it if the total numbers at different end digits of age were all compared to a tenth of the total population starting with age '0'. But this difficulty will be circumvented if for each different end digit of age, the population starting with that particular digit was only taken into account: thus, for digit '0' the proportion of the number returned with end digit '0' ages to the total population aged zero and above, for digit '1' the proportion of the number returned with end digit '1' ages to the population aged one and above, and so on, be taken. The proportions for different digits will not usually add up to unity under this method, but when reduced to a unit base the proportions should give proper relative measure of concentration for each of the individual digits.

6.4. The digit '0' may still get a slight weightage under this method owing to the incidence of the high infant mortality; but that will not be material; and the practical consideration is there that under-reporting of infants, a common feature of censuses and surveys, tends to offset this.

<sup>6.1</sup> *Journal of the Institute of Actuaries*, Vol. XLIX, p. 301.

<sup>6.2</sup> *Transactions of the Actuarial Society of America*, Vol. XLI, Part 2, No. 104, pp. 402-415.



# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

6.5. The measure of concentration suggested above and the index of aggregate distortion derived from it, are defined below in algebraic symbols. If  $n_{10v+u}$  denote the number actually returned at age  $10v+u$  then ' $m_u$ ' the measure of concentration at digit ' $u$ ' is given by

$$m_u = \frac{\sum_{v=0}^9 n_{10v+u}}{T_u} \times 10, \text{ where } T_u = \sum_{x=u}^9 n_x + \sum_{v=1}^9 \sum_{u=0}^9 n_{10v+u}.$$

And ' $I$ ' the index of aggregate distortion is given by

$$I = \sum_{u=0}^9 |m_u - 1|$$

It will be clear that ' $I$ ' will tend to zero in ideal conditions, if there were no distortion.

6.6 The measure of concentration for each digit and the index of aggregate distortion of the NSS medium for rural India, urban India and the city of Calcutta are given in Table (6.1). The measure and index for the population aged 40-above in the rural sector, for males and females separately in the urban sector (where the sex differential was found highest), and for the population of household-income group Rs.100 and below per month in the city of Calcutta have been actually presented in the table, to convey an idea as to how the concentrations vary from one population segment to another. The measures and index for Census 1951 in Uttar Pradesh individual age returns are also shown for comparison.

TABLE (6.1): MEASURES OF CONCENTRATION AT INDIVIDUAL END-DIGITS AND INDEX OF AGGREGATE DISTORTION IN AGE RETURNS

(NSS 4th round 1952, Calcutta Employment Survey 1953, and Census of India 1951)

(NSS 4th round 1952, Calcutta Employment Survey)						Calcutta Employment Survey (1,056 households)		Census 1951 <sup>6.4</sup> U.P. (males)
end digit	rural (28,918 persons)		urban <sup>6.3</sup> (28,715 persons)			all income groups	household income ≤ Rs. 100 per month	
	all ages	aged 40-above	persons	male	female			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	1.8	3.1	1.6	1.6	1.7	1.9	2.5	1.9
1	0.6	0.4	0.7	0.7	0.7	0.6	0.6	0.7
2	1.1	0.9	1.1	1.2	1.1	1.2	1.1	1.1
3	0.8	0.5	0.8	0.8	0.8	0.8	0.6	0.7
4	0.8	0.5	0.9	0.9	0.8	0.9	0.7	0.8
5	1.6	2.3	1.4	1.4	1.5	1.5	1.7	1.7
6	0.9	0.6	0.9	0.9	0.9	0.8	0.8	0.9
7	0.7	0.5	0.8	0.8	0.8	0.7	0.6	0.6
8	1.1	0.8	1.1	1.0	1.1	1.1	1.0	1.0
9	0.6	0.4	0.7	0.7	0.6	0.5	0.4	0.6
total	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
index of aggregate distortion	3.2	6.8	2.4	2.4	2.8	3.4	4.6	3.4

<sup>6.3</sup> Twelve out of total sixteen part-samples of NSS 4th round were used for this note.  
<sup>6.4</sup> Census of India 1951, Paper No. 3, 1954, pp. 36-37.



6.7. It will be seen that the urban pattern of concentration differed from the rural pattern on the one hand and the city pattern on the other. The most efficient set of grouping for the urban age returns need not therefore be the most efficient for the rural and the city sectors. The fact that not only the pattern of concentration of the household-income group Rs. 100 and below per month of the city of Calcutta was of different nature from the general city pattern, but that the distortion in their age reporting was even greater than in the general rural population, is somewhat unexpected. This apparently results from the break-up of the family and the drift from original community moorings of individuals in the lower income group; and the consequent failure of the applicability of a relative seniority ranking scale within the household and the community. The higher average age of the city population particularly in the lower income level, could also be a contributory cause. The implications of these differentials will be dealt with further in the next section.

6.8. It is easy to see from the distribution of numbers returned at individual ages (the diagrammatic representation of Chart (1) for example) that the force of concentration is comparatively low in the young age range; as could have been anticipated on *a priori* grounds, it gradually increases with advancing age. In the beginning, pulls of even over odd is most effective; then the pulls of '4' and '6' fade out giving place to increased pull for the middle digit '5'. Pulls of '0' and '5' dominate the middle age range, with '0' building up as age advances further. The position is complicated by existence of special pulls of the nature of age bias.

6.9. While the mounting nature of the pull of concentration has been noticed by earlier operators, no relative measures of deviation for the different age ranges appear to have been used so far. The root mean square deviation in decennial age ranges, calculated as the square-root of the sum of the squared deviations between the numbers actually returned and the expected frequencies, can provide such measures<sup>6.5</sup>. The difficulty which perhaps weighed with the operators in this field was that of estimating the expected frequencies. But simple assumptions like linear fall in expected frequencies between quinary pivotal values (estimated from the numbers actually returned by suitable grouping) should serve the purpose. The set of expected frequencies produced by even such crude assumptions would take account of the general shape of the true distribution and thus give satisfactory relative measures. The measures of concentration, taken along with such relative range measures of deviation could only give a proper insight into the extent and spread of the distortion.

<sup>6.5</sup> A number of other alternative measures of deviation are of course possible: one could be

$$\sum_{u=0}^9 \frac{(n_{10v+u} - r_{10v+u})^2}{r_{10v+u}}, \text{ giving the } \chi^2\text{-analogue of the distribution in the decennial age range } 10v \text{ to } 10v+9.$$



# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

6.10. In algebraic symbols,  $i_v$  the relative range measure of deviation of the decennial age range  $10v$  to  $10v+9$  is defined as,

$$i_v = \frac{\sqrt{\sum_{u=0}^9 (n_{10v+u} - r_{10v+u})^2}}{\sum_{u=0}^9 r_{10v+u}}$$

where  $r_{10v+u}$  is the expected number at age  $10v+u$ .

The expected number  $r_{10v+p}$  at pivotal age  $10v+p$ , ( $p = 4, 9$  for the 2:7 grouping adopted), was taken as

$$r_{10v+p} = \frac{1}{5} \sum_{u=p-2}^{p+2} r_{10v+u} ; \text{ and the expected numbers at other age,}$$

$$r_{10v+u} = r_{10v+p} - \frac{u-p}{5} (r_{10v+p} - r_{10v+p+5}), \text{ where, } u = p+1, p+2, \dots, p+4.$$

6.11. The relative range measures of deviation in the successive decennial age ranges for the population returned at individual ages in NSS for rural sector are given in Table (6.2). The progressive increase in the measures of deviation with advance of age is clearly brought out in columns (2) and (3) of the table. The root mean square deviation per cent of age (obtained by dividing the deviation per individual by the middle age of the range for simplicity and multiplied by 100) are also shown in the table : the interesting fact that the average deviation per unit of age attained is nearly uniform in all age ranges emerges from this part of the table.

TABLE (6.2) : RELATIVE RANGE MEASURES OF DEVIATION IN DECENNIAL AGE RANGES

(NSS 4th round 1952, all-India rural samples : 28,918 persons)

age range	deviation per individual		deviation percent of age	
	male	female	male	female
(1)	(2)	(3)	(4)	(5)
1. 0—9	0.03	0.02	0.61	0.43
2. 10—19	0.09	0.09	0.56	0.58
3. 20—29	0.15	0.15	0.61	0.59
4. 30—39	0.25	0.19	0.70	0.54
5. 40—49	0.29	0.25	0.63	0.56
6. 50—59	0.30	0.33	0.54	0.60
7. 60—69	0.37	0.50	0.57	0.76
8. 70—79	0.42	0.41	0.57	0.54
9. 80—89	0.36	0.58	0.42	0.69
10. 90—99	0.56	0.48	0.59	0.51



## SECTION SEVEN

## GROUPING EFFICIENCY

7.1. The efficiency of a set of grouping is traditionally assessed by the difference between the sum of the actual concentrations at the end digits comprised in the group and the ideal values.  $E_{0:5}$  the efficiency index of the quinary set of grouping 0 : 5 for example is given by  $\sum_0^4 m_x - 5$  : it is obvious that the complementary value  $\sum_5^9 m_x - 5$  will be the same, with just the sign reversed. This method was not altogether satisfactory in that the weight of numbers lay in the age range below 20 in population formations like India, and the group efficiency was accordingly determined to a greater extent by the pattern of concentration in this young age range : the chosen set of groups are however used throughout the span of life, and also for different socio-economic segments of the population, where the patterns of concentration are different.

7.2. It has been seen earlier how the distortions are comparatively small in the lower age range up to 20 and gradually increase with age. The relative efficiency of a set of grouping in the higher age ranges should thus provide a better indicator. It was therefore proper to decide on the most efficient set of grouping from a study of the behaviour of the various possible sets in different age ranges; the study should preferably extend to other socio-economic segments of the population. Table (7.1) gives the efficiency indices of the various possible sets of grouping in some important population segments.

TABLE (7.1): GROUP EFFICIENCY INDEX OF DIFFERENT SETS OF GROUPING  
(NSS 4th round 1952, Calcutta Employment Survey 1953, and Census 1951)

set of grouping	NSS 4th round : all-India					Calcutta Employment Survey (1,056 households)		Census 1951 U.P. (males)
	rural (28,918 persons)			urban (28,715 persons)		all income groups	household income ≤ Rs. 100 p.m.	
	all ages	aged 30-above	aged 40-above	male	female			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. 0 : 5	0.15	0.31	0.43	0.14	0.11	0.35	0.46	0.14
2. 1 : 6	-0.10	-0.36	-0.44	-0.01	-0.16	-0.07	-0.27	-0.05
3. 2 : 7	0.20	-0.06	-0.24	0.21	0.07	0.17	-0.01	0.16
4. 3 : 8	-0.24	-0.52	-0.62	-0.21	-0.27	-0.26	-0.49	-0.26
5. 4 : 9	-0.08	-0.18	-0.31	0.08	0.09	0.02	-0.11	0.01



## NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

7.3. Table (7.1) demonstrates how the relative variations in the group efficiencies, spread out in higher age segments of the same population. The urban and Calcutta samples were not analysed further in age segments, but the Calcutta sample was examined in an economic segment; in the household income  $\leq$  Rs. 100 p.m. population segment the variation again scattered wider. The group efficiency indices of the Census 1951 U.P. individual age distribution (1% sample)<sup>7.1</sup> is also shown in the table for comparison.

7.4. The rural sector is by far the more important; but on the basis of all ages there was not much to choose between the different sets of grouping, though the 4:9 and 1:6 sets had low indices. It has been pointed out earlier why the all ages index was not satisfactory. The 2:7 set shows the definite minimum indices at the higher age segments: it similarly has a definite minimum index in the lower income group, which again was by far the more important economic segment. The real test of efficiency is thus satisfied by the 2:7 set for which the index remains more stable and comparatively low in all the different segments, particularly where the indices for the other sets soar up: the index for this set also shows more changes in sign in passing through the population segments, which make for more balanced distribution of the group errors between it and its complementary group.

7.5. The 2:7 set of grouping was therefore adopted in analysis, interpretation and presentation of NSS data on age (as well duration). With a general tendency to over-estimate, the ages with the end digits '0' and '5' were likely to draw comparatively more from the ages below and the 2:7 set of grouping with the maximum concentration digits '0' and '5' placed towards the end of the groups, was also efficiently constituted to take account of this aspect of the estimation error.

7.6. Though the question of the most efficient set of grouping for the Indian Census was not in issue, examination and some discussion of the efficiency of grouping in the Census medium became inevitable. The mediums of collection of information and the types of errors are different for the Census and the NSS; the population is, however, the same and the relative efficiencies of the various sets of grouping could at least be expected to remain undisturbed as between them.

7.7. In Census 1931 Report<sup>7.2</sup> the 2:7 grouping was recommended after analysis of the age data in individual years on traditional lines. No detailed examination was done of Census 1941 age data. Numbers returned at individual ages in Census 1951 were not available for all India. Analysis of concentration and of group efficiency of the Census age material for Uttar Pradesh (U.P.), the only State which constitutes a Census population zone by itself, is shown in Table (6.1) and (7.1): in Census 1951 Age Tables<sup>7.3</sup> some detailed examination of the age data of the same

---

<sup>7.1</sup> *Census of India 1951, Paper No. 3, 1954, pp. 36-37.*

<sup>7.2</sup> *Census of India 1931, Vol. I, Part I, p. 135.*

<sup>7.3</sup> *Census of India 1951, Paper No. 3, 1954, p. 22.*



State was done to decide about the most efficient set of grouping. In the Age Tables the 2 : 7 set has been described as the standard grouping, but the 3 : 8 set has been recommended in the same breath as the 'proper' set; the belief that the dominant digits '0' and '5' should draw nearly equally from either side apparently tipped the scales in favour of the 3 : 8 set.

7.8. Primary grouping of the Census age returns in the 3 : 8 set however produced a saw-tooth distribution and a roller-type formula had to be applied to these quinary group values to get a smooth run. This was achieved by taking the weighted average of the group frequency concerned and the two adjoining group frequencies. The smooth set of group frequencies ultimately operated on for graduation purposes in Census 1951 thus rested on the assumption that the dominant digits drew from 7 individual ages on either side : such assumption looks stretched on the face of it. Table (7.1) showed clearly how the 3 : 8 set is the least efficient for the Indian situation. The problem of grouping efficiency exists even behind the operation of graduation : good graduation can only flow from an efficient set of group pivotal values.

7.9. It is interesting to note that if the numbers returned in individual ages in Census 1951 are grouped under the 2 : 7 set, the total deviations of the actual group frequencies from the expected (built up from the corresponding graduated individual frequencies) are smaller than similar total deviations of actual from expected under the 3 : 8 set of grouping; this was actually verified for the individual age distributions of Uttar Pradesh and Madras, special notice of which was taken in the Census 1951 Age Tables<sup>7.4</sup> to select the 'proper' grouping. When it is realised that the graduation itself is based on the 3 : 8 set, greater confidence is gained about the superiority of the 2 : 7 set.

7.10. Table (7.2) gives the comparative deviations of the actual numbers returned from the expected for the two competing sets of grouping 2 : 7 and 3 : 8 for the Census 1951 (males) population of Uttar Pradesh and Madras<sup>7.4</sup>.

7.11. The problem of grouping has been dealt so far with reference to age returns in individual years. But there may be situations when collection in individual years is either not possible or advised. The considerations guiding the selection of the most efficient groups will be altogether different if ages are as a rule not known or cannot be estimated in individual years. An ingenious suggestion made by R. Bachi (1951) to meet a somewhat similar situation deserves special mention in this context. He advised that all series which could not be collected by individual years might be collected for individual end digits '0' and '5' only, and for part-groups of end digits 1-4 and 6-9, as dominant distortions will be disclosed thereby<sup>7.5</sup>. Bachi further suggests routine methods of allocating the numbers returned at each of the preferred end digits '0' and '5', to the two bordering part-groups :

<sup>7.4</sup> *Census of India 1951, Paper No. 3, 1954, pp. 36-37 and 68-69.*

<sup>7.5</sup> *Bulletin of the International Statistical Institute, 1951, Vol. XXXIII, Part IV, pp. 218-221.*



# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

allocation in equal parts, or in proportion to the weights of the part-groups, or in inverse proportion to the relative broad aggregate measures of concentration of the part-groups (with adjustments for declining numbers with age) are the alternatives proposed.

TABLE (7.2): COMPARATIVE DEVIATIONS BETWEEN CENSUS NUMBERS RETURNED AND EXPECTED UNDER DIFFERENT SETS OF GROUPING

grouping set 2 : 7				grouping set 3 : 8			
age group (years)	number returned (000)	number expected (000)	deviation (2) — (3) (000)	age group (years)	number returned (000)	number expected (000)	deviation (2) — (3) (000)
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Uttar Pradesh (males) : Census 1951							
			+				+
			—				—
1. 2—6	4287	4191	96	3—7	4244	4241	3
2. 7—11	4129	4065	64	8—12	4453	3971	482
3. 12—16	3821	3606	215	13—17	3151	3505	354
4. 17—21	2744	3154	410	18—22	2981	3071	90
5. 22—26	2962	2806	156	23—27	2631	2745	114
6. 27—31	2476	2551	75	28—32	2709	2497	212
7. 32—36	2417	2293	124	33—37	2075	2238	163
8. 37—41	1986	2026	40	38—42	2114	1968	146
9. 42—46	1737	1764	27	43—47	1553	1710	157
10. 47—51	1538	1490	48	48—52	1609	1427	182
11. 52—56	1084	1182	98	53—57	961	1117	156
12. 57—61	908	871	37	58—62	922	809	113
13. 62—66	492	582	90	63—67	426	530	104
14. total	30581	30581	740	total	29829	29829	1138
average percentage deviation			4.84	average percentage deviation			7.63
Madras (males) : Census 1951							
			+				+
			—				—
1. 2—6	3701	3644	57	3—7	3607	3621	14
2. 7—11	3279	3398	119	8—12	3618	3339	279
3. 12—16	3541	3155	386	13—17	2949	3093	144
4. 17—21	2500	2806	306	18—22	2613	2724	111
5. 22—26	2500	2433	67	23—27	2326	2366	40
6. 27—31	2172	2165	7	28—32	2233	2118	115
7. 32—36	1951	1968	17	33—37	1793	1927	134
8. 37—41	1823	1774	49	38—42	1885	1730	155
9. 42—46	1482	1555	73	43—47	1360	1506	146
10. 47—51	1405	1320	85	48—52	1429	1269	160
11. 52—56	946	1062	116	53—57	857	1007	150
12. 57—61	862	804	58	58—62	871	751	120
13. 62—66	470	548	78	63—67	409	499	90
14. total	26632	26632	709	total	25950	25950	829
average percentage deviation			5.32	average percentage deviation			6.39

7.12. But the alternative allocations suggested by Bachi did not take account of the bias to over-estimate. Collection of age returns in individual years is possible in the situation contemplated by him, and small sample analysis does not involve much extra cost or time. A sample of the population (or of the Census slips) can indicate the estimation and other biases and should yield sufficiently accurate estimates of concentration and group efficiency. The most efficient set of grouping could be determined in this manner, in advance of the general tabulation, which may be done straightaway after that in the set of grouping adjudged most efficient.



APPENDIX 0

PROFORMA OF SCHEDULE USED IN THE ESTIMATES AND ESP STUDY

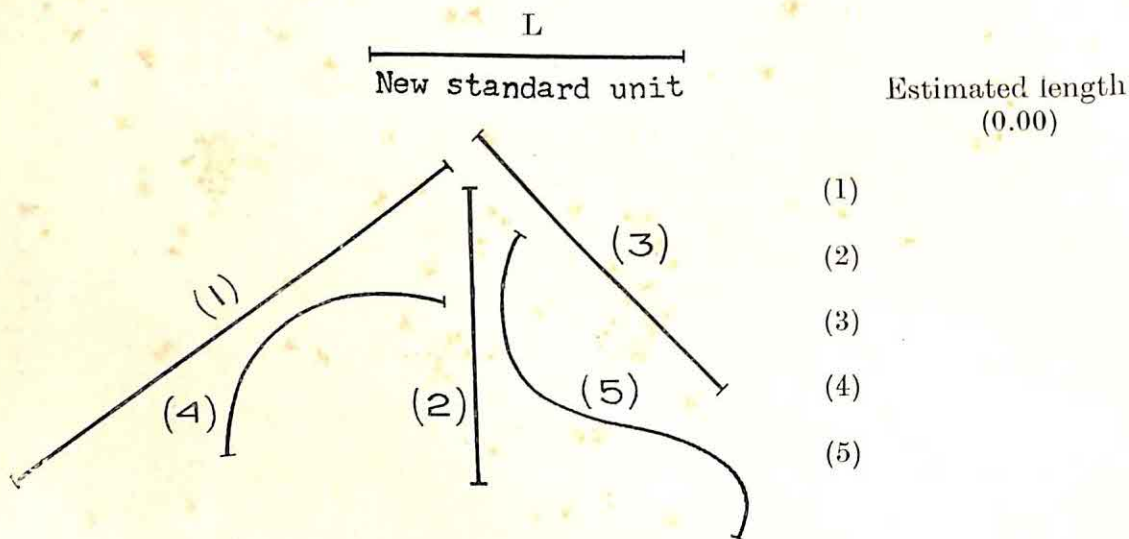
ESTIMATION AND ESP STUDY

The Demography Unit will be very much obliged if you kindly fill up the particulars below in your spare time and return the sheet to the Unit (Sri Samarendra Nath Mitra) early.

Roll No.....

Date.....


I. In terms of 'L' a new unit of linear measurement specified below, eye-estimate the lengths of the following five lines to two places of decimals and record the estimates :



II. The middle 3 digits of the following seven digitd numbers are missing : please complete the numbers by filling up the middle blank space with the digits that you think, on your first guess, might have been there.

(1)	93	85
(2)	27	47
(3)	45	90
(4)	12	08
(5)	66	19

Thank you !

  
 19 6 54  
 (Ajit Das Gupta)

# NATIONAL SAMPLE SURVEY : NOTES ON AGE GROUPING

## APPENDIX I

### DETAILED TABLES

TABLE 1(1): DISTRIBUTION OF INDIVIDUALS IN AGE-ASSESSED MINUS AGE-STATED CLASSES BY RELIGION

(NSS, WBSD Study 1954)

age-assessed minus age-stated	city (170 households)			other urban (405 households)			rural (754 households)		
	total	Hindu	others	total	Hindu	others	total	Hindu	others
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. age-assessed = age-stated (%)	495 (83.9)	422 (86.5)	73 (71.6)	1552 (86.2)	1296 (87.0)	256 (82.3)	2733 (77.6)	2160 (80.4)	573 (68.5)
2. age-assessed < age-stated (%)	20 (3.4)	15 (3.1)	5 (4.9)	70 (3.9)	53 (3.6)	17 (5.5)	274 (7.8)	184 (6.9)	90 (10.8)
3. age-assessed > age-stated (%)	75 (12.7)	51 (10.4)	24 (23.5)	178 (9.9)	140 (9.4)	38 (12.2)	515 (14.6)	342 (12.7)	173 (20.7)
4. total (%)	590 (100.0)	488 (100.0)	102 (100.0)	1800 (100.0)	1489 (100.0)	311 (100.0)	3522 (100.0)	2686 (100.0)	836 (100.0)

TABLE 1(2): DISTRIBUTION OF INDIVIDUALS IN AGE-ASSESSED MINUS AGE-STATED CLASSES BY MOTHER TONGUE

(NSS, WBSD Study 1954)

age-assessed minus age-stated	city (170 households)			other urban (405 households)			rural (754 households)		
	Bengali	Hindi	others	Bengali	Hindi	others	Bengali	Hindi	others
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. age-assessed = age-stated (%)	370 (84.9)	72 (75.8)	53 (89.8)	1120 (91.2)	261 (68.9)	171 (88.6)	2598 (78.1)	28 (70.0)	107 (68.1)
2. age-assessed < age-stated (%)	15 (3.4)	4 (4.2)	1 (1.7)	38 (3.1)	27 (7.1)	5 (2.6)	248 (7.5)	5 (12.5)	21 (13.4)
3. age-assessed > age-stated (%)	51 (11.7)	19 (20.0)	5 (8.5)	70 (5.7)	91 (24.0)	17 (8.8)	479 (14.4)	7 (17.5)	29 (18.5)
4. total (%)	436 (100.0)	95 (100.0)	59 (100.0)	1228 (100.0)	379 (100.0)	193 (100.0)	3325 (100.0)	40 (100.0)	157 (100.0)



TABLE 1(3): FREQUENCY DISTRIBUTION OF THE NUMBER RETURNED AT EACH INDIVIDUAL AGE BY SEX

(NSS 4th round 1952, all-India urban sample)

age (years)	male	female	persons	age (years)	male	female	persons
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
0	394	372	766	45	319	252	571
1	368	389	757	46	77	65	142
2	460	389	849	47	71	52	123
3	378	375	753	48	128	101	229
4	391	348	739	49	75	47	122
5	361	350	711	50	250	327	577
6	350	359	709	51	54	38	92
7	346	337	683	52	132	85	217
8	349	367	716	53	49	49	98
9	337	289	626	54	64	38	102
10	393	395	788	55	170	189	359
11	283	287	570	56	60	70	130
12	436	416	852	57	48	37	85
13	283	263	546	58	53	48	101
14	358	326	684	59	25	17	42
15	297	283	580	60	203	207	410
16	353	319	672	61	20	31	51
17	246	241	487	62	41	44	85
18	391	396	787	63	23	18	41
19	222	201	423	64	28	21	49
20	399	445	844	65	84	110	194
21	223	161	384	66	19	15	34
22	353	349	702	67	27	20	47
23	224	149	373	68	16	22	38
24	261	216	477	69	12	8	20
25	401	418	819	70	80	100	180
26	261	227	488	71	7	8	15
27	195	168	363	72	17	12	29
28	288	254	542	73	11	15	26
29	118	123	241	74	9	5	14
30	487	429	916	75	28	40	68
31	123	98	221	76	2	7	9
32	276	202	478	77	6	1	7
33	135	101	236	78	4	5	9
34	150	95	245	79	1	5	6
35	410	332	742	80	20	39	59
36	155	155	310	81	4	2	6
37	115	108	223	82	6	7	13
38	174	149	323	83	—	1	1
39	88	64	152	84	1	6	7
40	378	378	756	85	5	6	11
41	82	51	133	86	—	1	1
42	149	121	270	87	—	—	—
43	86	73	159	88	1	—	1
44	81	88	169	89	2	—	2

# MISCELLANEOUS

## A PARTIAL ORDER AND ITS APPLICATIONS TO PROBABILITY THEORY

By T. V. NARAYANA

*Institut H. Poincaré and McGill University*

**SUMMARY.** Application of a result in combinatory analysis, which generalizes the "problème du scrutin" of D. André to a partial order defined on the partitions of an integer. Relations of this partial order with two problems in probability theory.

### 1. DEFINITION OF PARTITION OF $n$

Given an integer  $n$ , we define an  $r$ -partition of  $n$  as follows:

An  $r$ -partition of  $n$ ,  $(t_1, \dots, t_r)$ , is a set of  $t_i$  where  $t_i \geq 1$  is an integer for  $i = 1, \dots, r$  such that

$$t_1 + \dots + t_r = n.$$

We remark that, in general, we shall consider  $(t_1, t_2, \dots, t_r)$ ,  $(t_2, t_1, \dots, t_r)$ , where  $t_1 + t_2 + \dots + t_r = n$ , as distinct  $r$ -partitions of  $n$ , unless  $t_1 = t_2$ . If  $r$  is an integer such that  $1 \leq r \leq n$ , we have, obviously,  $\binom{n-1}{r-1}$  distinct  $r$ -partitions of  $n$ .

### 2. PARTIAL ORDERING OF THE $r$ -PARTITIONS OF $n$

Given any two integers  $n, r$  ( $1 \leq r \leq n$ ), let us consider all the  $\binom{n-1}{r-1}$   $r$ -partitions of  $n$ . We shall say that an  $r$ -partition of  $n$

$$(t_1, \dots, t_r)$$

"dominates" the  $r$ -partition of  $n$

$$(t'_1, \dots, t'_r),$$

if and only if

$$t_1 \geq t'_1$$

$$t_1 + t_2 \geq t'_1 + t'_2$$

$$t_1 + \dots + t_{r-1} \geq t'_1 + \dots + t'_{r-1}.$$

... (2.1)

$$t_1 + \dots + t_r = t'_1 + \dots + t'_r = n.$$

Evidently

The relation of domination defined by (2.1) is reflexive, transitive and anti-symmetric. It thus represents a partial ordering of the  $r$ -partitions of  $n$ .

More generally, if  $(t_1, \dots, t_r)$  is an  $r$ -partition of  $m$ , and  $(t'_1, \dots, t'_r)$  is an  $r$ -partition of  $n$ , where  $m > n$ , we say that  $(t_1, \dots, t_r)$  dominates  $(t'_1, \dots, t'_r)$  if the relations (2.1) are satisfied.

Let us suppose we number the  $\binom{n-1}{r-1}$   $r$ -partitions of  $n$ , taken in some order, using the symbols  $p_1, p_2, \dots, p_{\binom{n-1}{r-1}}$ . Taking the partition  $p_i$ , let  $x_i$  be the number of partitions



dominated by  $p_i$  in the set  $p_1, p_2, \dots, p_{\binom{n-1}{r-1}}$ ;  $i = 1, \dots, \binom{n-1}{r-1}$ . The total  $(n, r) = x_1 + \dots + x_{\binom{n-1}{r-1}}$  obviously does not depend upon the particular ordering chosen for numbering the  $r$ -partitions of  $n$ . We state, as Lemma 1, the following result:

Lemma 1: For all integers  $n, r$

$$(n, r) = \binom{n}{r} \binom{n}{r-1} \dots \binom{n}{1} = n.$$

We shall demonstrate Lemma 1 in section 3 using a geometric interpretation of the  $r$ -partitions of  $n$ .

### 3. A COMBINATORIAL PROBLEM

Suppose we have a particle at the origin of an Euclidean space of  $k$  dimensions ( $k$  finite) and consider  $k$  mutually perpendicular axis. We shall be interested in points like  $p(a_1, a_2, \dots, a_k)$  where  $a_1 \geq a_2 \geq \dots \geq a_k \geq 1$  are all integers. We shall suppose that the particle can move on the network consisting of the points  $p$  following the rules given below :

Let  $a_{i\alpha}$  be the increase in the  $i$ -th coordinate at the  $\alpha$ -th step.

1)  $a_{i\alpha} \geq 1$  for all  $i, \alpha$  ( $i = 1, \dots, k; \alpha \geq 1$ ), i.e., at each step and following each axis the particle moves at least one unit.

2)  $a_{11} \geq a_{21} \geq \dots \geq a_{k1} \geq 1$ ;

$a_{11} + a_{12} \geq a_{21} + a_{22} \geq \dots \geq a_{k1} + a_{k2} \geq 2$

and, in general, for the  $\alpha$ -th step

$$\sum_{j=1}^{\alpha} a_{ij} \geq \sum_{j=1}^{\alpha} a_{i'j} (\geq \alpha) \text{ when } i \leq i', i' = 1, \dots, k.$$

Let us suppose that we know that at the  $r$ -th step the particle has reached the point  $(a_1, a_2, \dots, a_k)$ ,  $a_1 \geq a_2 \geq \dots \geq a_k \geq r$ , and let  $(a_1, \dots, a_k)_r$  be the total number of different ways by which the particle can arrive at this point (or the total number of possible paths).

Theorem 1: We have :

$$(a_1, \dots, a_k)_r = \begin{vmatrix} (a_1-1)_{(r-1)} & (a_2-1)_{(r)} & \dots & (a_k-1)_{(r+k-2)} \\ (a_1-1)_{(r-2)} & (a_2-1)_{(r-1)} & \dots & (a_k-1)_{(r+k-3)} \\ \dots & \dots & \dots & \dots \\ (a_1-1)_{(r-k)} & (a_2-1)_{(r-k+1)} & \dots & (a_k-1)_{(r-1)} \end{vmatrix}$$

where

$$(a_i-1)_{(t)} = (a_i-1)_t.$$

This theorem generalizes the André-Poincaré "problème du scrutin" H. Poincaré (1913), E. Borel (1925).

# A PARTIAL ORDER AND ITS APPLICATIONS TO PROBABILITY THEORY

*Demonstration :* We shall prove the theorem for  $k = 2$ . (The general proof will be given as a special case in a forthcoming paper of the author). The network now consists of the points  $p(a_1, a_2)$  where  $a_1 \geq a_2 \geq 1$ .

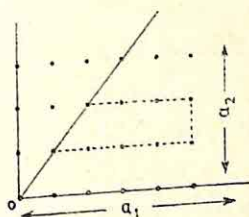
Evidently

$$\begin{aligned} (a_1, a_2)_1 &= 1 \quad \text{for } a_1 \geq a_2 \geq 1 \\ (a_1, a_2)_2 &= 0 \quad \text{if } a_2 < 2 \\ &\quad \text{or } a_1 < a_2. \end{aligned}$$

If  $a_1 \geq a_2 \geq 2$ ,  
 $(a_1, a_2)_2 = \sum \sum (a_1, a_2)_1$ , where the summation extends over the points  $p$  contained in the truncated rectangle formed by the origin and the point  $(a_1, a_2)$  as shown.

Thus

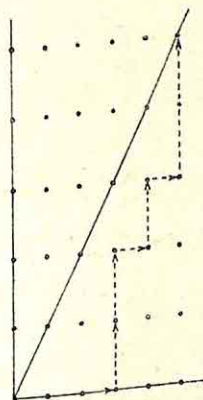
$$\begin{aligned} (a_1, a_2)_2 &= \sum_{a_1 \geq a_2} \sum_{a_2=1}^{a_2-1} 1 \\ &= (a_1-1)(a_2-1) - (a_2-1)_{(2)}. \end{aligned}$$



We can prove easily, by induction, that :

$$(a_1, a_2)_r = (a_1-1)_{(r-1)}(a_2-1)_{(r-1)} - (a_1-1)_{(r-2)}(a_2-1)_{(r)}.$$

Set  $a_1 = a_2 = n$ .  
 Each path represents a domination of an  $r$ -partition of  $n$  by an  $r$ -partition of  $n$ ; and, inversely, to each domination of an  $r$ -partition of  $n$  by an  $r$ -partition of  $n$ , corresponds a path. We give below an example with  $n = 5$ ,  $r = 3$ .  $(3, 1, 1)$  dominates  $(2, 1, 2)$ .



We thus obtain the result of Lemma 1; for :

$$(n, r) = (a_1, a_2)_r \quad \text{when } a_1 = a_2 = n \geq r$$

or

$$(n, r) = \frac{\binom{n}{r} \binom{n-1}{r-1}}{n}, \quad (1 \leq r \leq n).$$

We solve—once again—the problème du Scrutin of Désiré André from this result.



4. AN EXTENSION OF THE "PROBLÈME DU SCRUTIN"

The 'problème du Scrutin' is as follows : "Two candidates  $A$  and  $B$  stand for an election. A well-informed observer knows beforehand that  $A$  will obtain  $m$  votes and  $B$   $n$  votes, where  $m > n$ . What is the probability that  $A$  will lead  $B$  throughout the scrutiny of the votes?"

Since  $A$  leads  $B$  throughout the scrutiny, equality of votes not being allowed, we might reword the problem as follows: What is the probability that  $A$  holds a 1-lead over  $B$  throughout the scrutiny? We solve below the more general question: What is the probability that  $A$  holds a  $L$ -lead over  $B$ ,  $L$  being an integer so that  $1 \leq L \leq m-n$ ?

*Solution* : Given  $1 \leq L \leq m-n-1$ ,  $A$  could hold the  $L$ -lead over  $B$  in the following mutually exclusive ways :

The last vote is for  $B$ .

The last vote is for  $A$ ; but the last but one is for  $B$ .

The last two votes are for  $A$ ; the preceding one is for  $B$ .

The last  $(m-L-n)$  votes are for  $A$ ; the preceding one is for  $B$ .

Let us consider the case where the last vote is for  $B$ . Evidently for  $A$  to hold the  $L$ -lead over  $B$ , the first  $L$ -votes (at least) have to be for  $A$ . We now remark that there is a one-to-one correspondence between each domination of an  $r$ -partition of  $n$  (the votes for  $B$ ) by an  $r$ -partition of  $m-L$  (the remaining votes for  $A$ ) and sequences of  $m$   $A$ 's and  $n$   $B$ 's where  $A$  holds the  $L$ -lead over  $B$ . Since

$$(m-L, n)_r = (m-L-1)_{(r-1)}(n-1)_{(r-1)} - (m-L-1)_{(r-2)}(n-1)_{(r)},$$

the number of ways in which  $A$  holds the  $L$ -lead over  $B$ , the last vote being for  $B$ , is

$$\sum_{r=1}^n (m-L, n)_r = (m+n-L-2)_{(n-1)} - (m+n-L-2)_{(n-3)}.$$

The case where the last vote is for  $A$ , (the last but one being for  $B$ ) gives rise, similarly, to

$$\sum_{r=1}^n (m-L-1, n)_r = (m+n-L-3)_{(n-1)} - (m+n-L-2)_{(n-3)}$$

ways in which  $A$  holds the  $L$ -lead over  $B$ .

If just the last  $(m-L-n)$  votes are for  $A$ , we have similarly

$$\sum_{r=1}^n (n, n)_r = (2n-2)_{(n-1)} - (2n-2)_{(n-3)}$$

ways in which  $A$  can hold the  $L$ -lead over  $B$ . The total number of ways in which  $A$  holds the  $L$ -lead over  $B$  is

$$(m+n-L-1)_{(n)} - (m+n-L-1)_{(n-2)}.$$

The probability of holding the  $L$ -lead is thus

$$\frac{m!(m+n-L)!(m-L-n+1)}{(m+n)!(m-L+1)!}$$

which reduces to  $\frac{m-n}{m+n}$  for  $L = 1$ .

# A PARTIAL ORDER AND ITS APPLICATIONS TO PROBABILITY THEORY

The probability of holding the  $L$ -lead and no better is evidently

$$\frac{m!(m+n-L-1)!n(m-L+2-n)}{(m+n)!(m-L+1)!}.$$

For  $L = m-n$ , it is easy to see that the number of ways in which  $A$  holds the  $L$ -lead over  $B$  is

$$(2n-2)_{(n-1)} - (2n-2)_{(n-3)}$$

a result independent of  $m$ .

We remark in passing that in order that  $A$  holds the 1-lead over  $B$  with probability one-half, it is necessary that  $m = 3n$ . A simple calculation shows that, when  $n$  is large,  $m$  should equal  $(2+\sqrt{5})n$  or  $4.3n$  approximately, if  $A$  should hold the 2-lead over  $B$  with the same probability.

## 5. A PROBLEM IN PROBABILITY THEORY

Let us suppose that we are given two coins 1, 2 with probabilities  $p_1, p_2$  of obtaining heads, and consequently the probabilities  $q_1, q_2$  of obtaining tails where  $q_i = 1-p_i$ ,  $i = 1, 2$ . We shall assume in what follows that  $p_1+p_2 > 1$ .

Let us consider the game  $G_2$  played with the following rules: Narayana (1955).

- 1) The first trial is made with coin 1.
- 2) For  $n > 1$ , the  $n$ -th trial is made with coin 1 or coin 2, according as the result of the  $(n-1)$ st trial was a tail or head.

- 3) We stop the series of trials at that trial where for the first time the accumulated number of heads obtained (with both coins) is greater than the accumulated number of tails obtained by exactly 2.

Since we have assumed  $p_1+p_2 > 1$ , the probability that our game  $G_2$  will terminate in a finite number of trials approaches unity as closely as we please.

It is evident that the game  $G_2$  can end only at the  $(2n+2)$ nd trial,  $n = 0, 1, 2, \dots$ , and that the last, i.e.,  $(2n+2)$ nd trial is made with coin 2. We shall use a simple, self-evident notation to represent a sequence of trials, letting  $x$  represent a head and  $o$  a tail. The sequence given below

$$\begin{array}{ccccccc} x & o & o & x & & & \\ & & & & x & x & x \\ & o & & & & & \end{array}$$

indicates that the game  $G_2$  is terminated at the 8th trial and that :

The 1st trial was a head with coin 1.

The 2nd trial was a tail with coin 2.

The 3rd trial was a tail with coin 1, and so on.

We shall say that a sequence of trials representing  $G_2$  belongs to the series  $S_n$ ,  $n \geq 0$  being integer or zero, if we obtain  $n$   $o$ 's with coin 1 during the sequence. For example, the sequence above, since it contains two tails with coin 1, (3rd and 4th trials), belongs to  $S_2$ . We thus



classify a sequence belonging to  $G_2$  according as the number of  $o$ 's obtained with coin 1 in this sequence, into mutually exclusive classes  $S_0, S_1, S_2, \dots, S_n, \dots$ . Any sequence of  $G_2$  necessarily belongs to one and only one of these series. We shall study the first few of these series to define the concept of "base" sequence.

*Series  $S_0$* : Since any sequence of  $G_2$  belonging to  $S_0$  contains no  $o$ 's with coin 1, it is clear after some consideration, that  $S_0$  contains sequences of the type:

$$\begin{array}{ccccccc} x & & x & x & & x & x & x \\ ; & & & ; & & & ; & \dots \\ x & & o & x & & o & o & x \end{array}$$

and in general, a sequence containing  $(2n+2)$  trials and belonging to  $S_0$  can occur only in one way, viz., :

$$\begin{array}{ccc} \leftarrow n \rightarrow & & \\ x & x & x \\ \dots & & \\ o & o & o \end{array}$$

i.e.,  $n$  patterns of the type  $x_0$  followed by  $x_x$  which terminates the event. We call  $x_x$  a "base" sequence for  $S_0$  for reasons to be obvious shortly. Any term of  $S_0$  can be obtained from the base sequence by adding an approximate number of "subsidiary patterns"  $x_0$  suitably, i.e., so as not to contradict the rules of  $G_2$ .

*Series  $S_1$* : A little consideration will show that there exists one and only one base sequence for  $S_1$  from which every term of  $S_1$  can be obtained by adding an appropriate number of "subsidiary patterns" of the type  $x_0$  or  $o_x$  suitably. The base sequence is given by:

$$\begin{array}{cc} o & x \\ & x & x \end{array}$$

and we can add  $x_0$  instead of any line indicated by  $\backslash$  and  $o_x$  instead of any line indicated by  $/$ , in the base sequence as shown below :

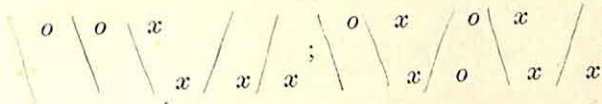
$$\begin{array}{ccc} \text{(i)} & \text{(ii)} & \text{(iii)} \\ \backslash & o & \backslash & x & \text{(iii)} \\ & & & x & / & x \end{array}$$

Thus, for example, if we would like to obtain all the sequences of  $G_2$  belonging to  $S_1$  containing 6 terms we need only add  $x_0$  or  $o_x$  suitably, giving the sequences

$$\begin{array}{lll} \text{(i)} & x & o & x \\ & o & & x & x ; \end{array} \quad \begin{array}{lll} \text{(ii)} & o & x & x \\ & & o & x & x ; \end{array} \quad \begin{array}{lll} \text{(iii)} & o & x & x \\ & & x & o & x . \end{array}$$

The problem is the same as that of putting  $\frac{6-4}{2} = 1$  ball in 3 boxes, the boxes being indicated by the sloping lines in the base sequence. In general, if a sequence of  $S_1$  contains  $(2n+4)$  terms,  $n = 0, 1, 2, \dots$ , this sequence can occur in the same number of ways as that of putting  $\frac{(2n+4)-4}{2} = n$  balls in 3 boxes or in  $\binom{n+2}{2}$  ways.

Series  $S_2$ : We shall state that there are 2 base sequences for the series  $S_2$ , viz.,



The number of "boxes" is 5, indicated by the sloping lines and if a sequence of  $S_2$  consists of  $(2n+6)$  terms,  $n = 0, 1, 2, \dots$ , this sequence can occur in

$$\binom{n+4}{4} + \binom{n+3}{4} \text{ ways.}$$

## 6. DEFINITION OF BASE SEQUENCES

Given the series  $S_n$ , i.e., the total number  $n$  of tails occurring with coin 1 during a sequence of trials of  $G_2$ , a set of base sequences for the series is a set of sequences of observations in  $S_n$ , from which all other sequences of the series can be obtained by inserting suitably any number of subsidiary patterns of the form  $x_0$  or  ${}_0x$  in the sequences. Any sequence in this set of base sequences will be called a base sequence. It can be shown easily that a base sequence of  $S_n$  ( $n \geq 1$ ) can contain either  $(2n+2)$  or  $(2n+4)$  or ...  $(4n)$  trials. For  $n \geq 1$ , let us define a base sequence of  $S_n$  of length  $r$ , as a sequence  $(2n+2r)$  trials ( $1 \leq r \leq n$ ). We state a theorem (which can be proved by simple methods) which gives us all the base sequences for  $S_n$  ( $n = 1, 2, \dots$ ) and shows the relation of the game  $G_2$  with the dominations of the partitions of an integer.

Theorem 2: To every domination of an  $r$ -partition of  $n$  by another there corresponds a base sequence of  $S_n$  of length  $r$  and conversely.

Thus  $(n, r) = \binom{n}{r} \binom{n}{r-1} / n$  represents the number of base sequences of  $S_n$  of length  $r$  and the total number of base sequences of  $S_n$  is

$$\sum_{r=1}^n (n, r) = \frac{1}{n+1} \binom{2n}{n}.$$

Making the convention  $(0, 1) = 1$ , i.e., the number zero dominates itself, the theorem is true for all integral  $n \geq 0$ .

Finally, since the probability, that the game  $G_2$  ends in a finite number  $n$  of trials, approaches unity as  $n \rightarrow \infty$ , we have the identity

$$\frac{p_1 p_2}{(1-p_1 q_2)} + \frac{q_1 p_1 p_2^2}{(1-p_1 q_2)^3} + \frac{q_1^2 p_1 p_2^3}{(1-p_1 q_2)^5} (1+p_1 q_2) + \frac{q_1^3 p_1 p_2^4}{(1-p_1 q_2)^7} (1+3p_1 q_2 + p_1^2 q_2^2) + \dots = 1,$$

where  $p_1 + p_2 > 1$ ,  $p_i = 1 - q_i$  ( $i = 1, 2$ ),

the general term of this series being :

$$\frac{q_1^r p_1 p_2^{r+1}}{(1-p_1 q_2)^{2r+1}} \frac{1}{r} \left[ \binom{r}{1} \binom{r}{0} + \binom{r}{2} \binom{r}{1} p_1 q_2 \dots + \binom{r}{r} \binom{r}{r-1} p_1^{r-1} q_2^{r-1} \right].$$



6, ACKNOWLEDGEMENTS

I am very grateful to Dr. N. L. Johnson for having suggested this problem and for his continued interest and aid in this piece of work. I would like also to express my thanks to Professor G. Darmon and Professor M. Loève for their kind interest and encouragement.

REFERENCES

- BOREL, E. (1925): *Traité de Calcul des Probabilités et ses Applications*. Paris, Gauthier Villars et Cie.  
 POINCARÉ, H. (1913): *Calcul des Probabilités*. Paris, Gauthier Villars et Cie.  
 NARAYANA, T. V. (1953): Sequential Procedures in Probit Analysis. *Doctoral thesis submitted to the University of North Carolina*.  
 ——— (1955): *Comptes Rendus*, t. 240, 1188-89.

*Paper received : May, 1957.*

# RANDOM PROCESSES IN ECONOMIC THEORY AND ANALYSIS

By P. A. P. MORAN

*Australian National University, Canberra*

**SUMMARY.** The various models of discrete parameter random processes used in econometrics and economic theory are reviewed, and a summary given of the known results in the theory of testing hypotheses and estimating parameters for such models. It is shown that for empirical economic series it may be difficult or impossible to make an adequate verification of the hypotheses on which such methods are based. The example of testing the correlation between two short series is considered in details. Finally an outline is given of the identification and estimation problem for multivariate processes.

## 1. THE USE OF MODELS IN ECONOMICS

Economic theory can only be either useful or substantiated when it has been related to empirical facts which are necessarily almost entirely of a numerical kind. There are so many of these facts that it is necessary to be able to summarize them or otherwise describe them in short form and thus some kind of descriptive statistical method is necessary in economics. Even when this is done, however, they rarely show clear patterns which can be easily interpreted in terms of economic theory. It is this fact which makes economics such a different science from physics. Moreover, unlike physics again, there are usually so many other influences, often unmeasurable, at work besides those considered in the theory that even when some kind of pattern appears to show itself, there can be no great assurance that this is not due to random causes not considered in the theory.

For these reasons it is necessary to go further and use theoretical statistics in an attempt to sort out which influences on the data are of a systematic kind and which can be regarded as, in some sense or other, random influences. Only in this way can one guard against the easy errors involved in arguing from observed patterns which are in fact only due to chance.

Theoretical statistics is necessarily based on probability theory and our method of procedure must therefore be to construct a theoretical probability model which, we hope, is a close representation of actual processes resulting in the observations used. This setting up of a model in which some or all of the variables considered are random variables, i.e., have associated with them probability distributions, is an essential part of the job of comparing an economic theory with empirical data. However economic models involving probability may also throw some light on economic theory. For example some dynamic economic models may be shown to be intrinsically unstable when small random influences are incorporated into the model. Again one plausible theory to explain why some economic quantities show a quasi-cyclic behaviour is that these variables are, in principle, determined by a set of equations determining a stable equilibrium but that this system is disturbed by random shocks. For certain values of the constants involved the system, although stable, may show a tendency to overshoot its equilibrium values and thus show a quasi-cyclic behaviour.

In setting up a model it is first of all necessary to be quite clear as to which quantities are to be supposed to have probability distributions. It is therefore convenient to distinguish



between quantities which are supposed to remain fixed in the model and quantities with a probability distribution. We shall call both kinds of quantities *variables* but those variables with a probability distribution we shall also call *variates*. The joint distribution of all the variates in the model may be called the probability set of reference. This is a conceptual set of possibilities which is used in order to draw inferences from the statistical data. The value and necessity of making quite clear what this set is in each particular case will be illustrated in the examples which follow.

The grounds for adopting a particular model are partly background knowledge of the situation and intuition into its structure, and partly based on examination of the actual data. Thus, for example, in a physical measurement of a length, past experience together with theoretical knowledge both suggest that the distribution of errors of measurement may be supposed to be the normal distribution. Further evidence may also be obtained by applying tests of normality to the observations, but, if these are few in number this evidence may be slight.

As will be shown by later examples, a false model may lead to totally incorrect conclusions. On the other hand, such a model may, in some circumstances, be quite useful, especially in prediction. For example in studying annual sunspot numbers a prediction based on a regression equation connecting each annual value with the two previous values will result in a fair predictor; but this is certainly not an adequate description of whatever process really underlies the system.

Having set up a probability model of a plausible kind we may proceed to test various hypotheses and to estimate the various parameters in the model. The setting up of a test of any hypothesis requires some care. The test must be relevant to the kind of divergence from the hypothesis which is in view and this requirement has led to the introduction of the idea of the power of a test. Moreover the test criterion may tend to show apparently significant results, not because the particular specific hypothesis is incorrect but because the whole model is not correct. Thus in testing whether the mean of a set of observations is significantly different from some specified value by using the *t*-test, apparently significant results may be obtained, not because the true mean is in fact different but because the distribution is not normal. This has led to the introduction by G.E.P. Box of the expression, the 'Robustness' of a test, which may be defined to be the insensitivity of the test to some particular type of divergence from hypothesis other than that to which the test is relevant.

Further difficulties arise when we turn to the problem of estimating parameters in the model. To simplify the situation as much as possible suppose the model involves parameters  $\theta_1, \dots, \theta_k$  and the observations are a sample  $x_1, \dots, x_n$  of a single variate which has a probability distribution  $f(x/\theta_1, \dots, \theta_k)$  specified by the model. From the sample we construct functions  $t_1(x_1, \dots, x_n), \dots, t_k(x_1, \dots, x_n)$  which we use as our estimators of  $\theta_1, \dots, \theta_k$ . We do not consider here the various well-known criteria which it would be desirable for such estimators to possess but ask the more fundamental question whether the  $\theta_i$  are in principle estimable at all. This is the problem of identifiability and may be illustrated by a simple example. Suppose that we set up a model to explain a set of observations  $\{x_i\}$  in the following way. We suppose, from our knowledge of the situation, that  $x = y + z$  where  $y$  and  $z$  are independent normal variates with means  $m_1, m_2$  and standard deviations  $\sigma_1$  and  $\sigma_2$ . Then we soon see that not even the introduction of Banach spaces will enable us to estimate these four quantities separately and they are, in principle, not estimable individually. They



are said to be unidentifiable. We can, however, estimate  $m_1 + m_2$  and  $\sigma_1^2 + \sigma_2^2$  and these quantities are therefore said to be identifiable. A practically more relevant example of non-identifiability is that of the constants in a pair of demand-supply relationships and a clear discussion of this is given by Koopmans (1949).

The problem of identification is not only of importance theoretically but also practically, for a policy decision in an economic situation may have an effect which depends on a parameter unidentifiable from the observations and the effect of the policy decision may therefore be unpredictable. Illustrations of this and other examples of non-identifiability will occur later.

It is rare to have a situation in which there is only one kind of measurement or variable. We usually have to deal with situations in which each element of the sample consists of two or more measurements. Before considering how random processes enter into our models, it is of value to consider the present situation of the theories of correlation and regression, a subject in which there has been much confusion. Let us suppose we are confronted with an empirical situation in which we have  $n$  pairs of values,  $(x_1, y_1), \dots, (x_n, y_n)$  and we are going to set up a probability model in order to make inferences from the data. As we are going to ignore, for the moment, all the problems which result from the use of random processes we shall always assume that whatever model we use, different pairs  $(x_i, y_i), (x_j, y_j)$  will be distributed independently. To investigate the relationship between  $x$  and  $y$  we can set up at least four different types of model.

Consider first the model of linear regression (non-linear regression models are of the same type and may be considered in the same way). Here we suppose that the  $x_i$  are simply mathematical quantities or variables with no probability distribution attached to them. We shall suppose the  $y_i$ , on the other hand, have a probability distribution whose mean is a linear function of  $x$ . It is usually satisfactory to assume further that this distribution is normal or Gaussian and that its standard deviation is a constant (nearly always unknown) independent of  $x$ . Notice that in this model all the probability is related to the  $y$  variable and so we say that in this model  $y$  is a variate and  $x$  only a mathematical variable. The probability set to which our inferences are related then consists of the joint distribution of all the  $y$ 's for the fixed observed set of  $x$ 's. Using this distribution we can set up various tests of significance and estimators. For example, if the mean value of  $y$  is taken to be a linear function,  $\alpha + \beta x$ , of  $x$  we can test the hypothesis that  $\beta = 0$  by the test criterion

$$t = \frac{bs_1\sqrt{(n-2)}}{\sqrt{(S_2^2 - b^2S_1^2)}}$$

which is distributed in Student's  $t$ -distribution with  $n-2$  degrees of freedom. Here

$$b = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

$$S_1^2 = \frac{1}{n-1} \sum(x_i - \bar{x})^2$$

$$S_2^2 = \frac{1}{n-1} \sum(y_i - \bar{y})^2.$$

The parameters in this model are  $\alpha$ ,  $\beta$  and the variance of  $y$  for fixed  $x$ . All these parameters are identifiable.



A model of completely different kind is obtained when both variables are assumed to be random variables, i.e., we assume that the pair  $(x_i, y_i)$  are jointly distributed in a bivariate probability distribution and we may ask first whether there is any evidence to show that  $x_i$  and  $y_i$  are not distributed independently. The natural tool to use in this case is the correlation coefficient

$$r = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2\}}}.$$

Under the hypothesis that  $x$  and  $y$  really are distributed independently, it is easy to show that the mean and variance of the distribution of  $r$  are zero and  $(n-1)^{-1}$  respectively. The exact form of the distribution depends on the distribution of  $x$  and  $y$  but tends to normality with increasing  $x$ . It is a remarkable fact that if we now assume merely that one of the two variates  $x$  and  $y$  is normally distributed, we can deduce the exact distribution which is given in the standard text books and is well tabulated, and which applies, a fortiori, when both  $x$  and  $y$  are normally distributed. It is known, moreover, that the null distribution of  $r$  is relatively insensitive to joint variation of the distributions of  $x$  and  $y$  from normality and in this respect the test based on  $r$  is said to be robust.

We also notice that, purely algebraically,

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

and that the distribution of  $r$ , when  $x$  and  $y$  are independent and one a normal variate, can be obtained from that of  $t$ . In the two cases we calculate the same quantity  $t$ , and under the assumption of the null hypothesis, ascribe to it the same distribution. The two cases are, however, fundamentally different because they refer to different probability sets or universes of reference. In the regression case we considered the universe of all possible values of  $y$  given the fixed observed values of  $x$ , whereas in the second case we considered a universe of reference in which the  $x$ 's also vary. The difference between the two tests becomes a matter of practical importance when we consider their power relative to specified alternatives, e.g., that  $\beta \neq 0$ .

Finally we notice that in the correlation model, the parameters of the population, the means, variances and correlation coefficient of  $x$  and  $y$  are all identifiable and can be efficiently estimated from a sample.

A model of more relevance in economic applications is that known as the 'Error in Variables Model'. Here we suppose the observations  $(x_i, y_i)$  to be random variables obtained as measurements or estimates of quantities  $u_i, v_i$  which are linearly related. To be specific we suppose that  $v_i = \alpha + \beta u_i$  ( $i = 1, \dots, n$ ) where  $\alpha$  and  $\beta$  are quantities describing the *structure* of the model and the  $u_i, v_i$  may be either fixed quantities or variates but are, in any case, unobserved. Next we suppose that

$$\begin{aligned} x_i &= u_i + \epsilon_i \\ y_i &= v_i + \eta_i, \end{aligned}$$

where the  $\epsilon_i, \eta_i$  are sets of random variables which are independent of each other for different  $i$  and which may, in most cases, be assumed to be independent. It will then be found that  $\alpha$  and  $\beta$  are not identifiable.



## RANDOM PROCESSES IN ECONOMIC THEORY AND ANALYSIS

To be more specific suppose that  $u_i$  is a normal variate with unknown mean  $m$  and unknown standard deviation  $\sigma_1$ .  $v_i$  is then a normal variate with mean  $\alpha + \beta m$  and standard deviation  $\beta\sigma_1$ . Suppose that  $\epsilon_i, \eta_i$  are normal variates with zero means and standard deviations  $\sigma_2$  and  $\sigma_3$ . Then it is obvious that  $x_i$  and  $y_i$  are normal variates with means  $m, \alpha + \beta m$ , standard deviations  $\sqrt{(\sigma_1^2 + \sigma_2^2)}, \sqrt{(\beta^2\sigma_1^2 + \sigma_3^2)}$  and correlation

$$\frac{\beta\sigma_1^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)}\sqrt{(\beta^2\sigma_1^2 + \sigma_3^2)}}.$$

It is easy to see that these five quantities are identifiable from the sample but even if they are known exactly, we could not deduce from them the values of  $\alpha$  and  $\beta$  which are thus unidentifiable.

The importance of this in economics arises from the fact that it may be important to know the values of  $\alpha$  and  $\beta$  rather than the observed regressions. For, if by some policy decision we could increase  $u_i$ , say, by some specified amount, we would like to know what would be the effect on  $E(y_i)$ , the expected value of  $y_i$ , and this is impossible if we do not know  $\beta$ .

Although we cannot estimate  $\beta$ , it is possible to place bounds on the underlying relationship between  $v$  and  $u$ . This linear relationship must pass through the true means of  $x$  and  $y$  (a point which can be estimated) and have a slope intermediate between the slopes of the lines giving the regression of  $y$  on  $x$  and  $x$  on  $y$ . More than this we cannot say without some further knowledge, such as, for example, the ratio of the variances  $\sigma_2^2$  and  $\sigma_3^2$ . It may in fact be useful to know that the true relationship lies between certain limits and this, in a more complicated context, is the idea underlying Frisch's bunch map analysis (see for example Stone<sup>1</sup> (1945, 1954)). Furthermore it is also possible to take account of the uncertainty of the sample values of the regression coefficients and set up a test for a specified value of  $\beta$  (Moran, 1956). This test is only useful in some circumstances in enabling us to *reject* some particular values of  $\beta$  and for large samples the region, in which rejection does not take place, does not vanish. Another method of getting around the present difficulties is the use of 'Instrumental Variables'. For this see Stone (1954) and Reiersol<sup>2</sup> (1945).

There is, finally, another model which is interesting in itself but not likely to be useful in economics. This is the Berkson model (Berkson, 1950; Lindley, 1953). Here we suppose that  $x_1, \dots, x_n$  are a set of previously specified fixed values such as might be chosen as a set of values at which some parameter in a physical experiment is to be fixed. However the actual experimental values cannot be prescribed exactly and are in fact  $z_1, \dots, z_n$  where  $z_i - x_i$  are random variables, the errors of specification, which are independent of each other and of the  $x$ 's. We then suppose that

$$y_i = \alpha + \beta z_i + \epsilon_i$$

where the  $\epsilon_i$  are further random variables independent of  $z_i$  and  $x_i$ . In this case it is easy to see that  $\alpha$  and  $\beta$  are identifiable. This is the situation which is likely to occur in physical experimentation but does not seem relevant in economics.

---

<sup>1</sup>Notice, however, that Stone calls the underlying relationship a "regression," and also, for example, takes the regression of  $x$  on  $y$ , inverts it and calls the resulting relationship a "regression." This is not the usual terminology.



In all the above models we have assumed that different pairs  $(x_i, y_i)$  are independent of each other and have the same distribution. Largely as a result of the writings of Yule (1926) it became realized that there are many cases in which this does not apply and that very misleading conclusions may result. Yule calculated the correlation between the standardized mortality per 1000 persons, and the proportion of Church of England marriages to total marriages for the years 1866-1911 and found  $r = 0.9512$ . This appears to be highly significant by the ordinary test. He rightly describes this as a 'nonsense correlation' and its origin is clearly due to the fact that both series of figures show a strong trend. In this case one can regard the nonsense correlation as being due to the fact that both variates are strongly dependent on a third variable, time, and the fallacy of ascribing a causal connection is thus an elementary and well-known one. A less obvious fallacy arises when neither series shows a trend but both are serially dependent. If we have two series, such as annual sunspot numbers and some economic variable which has no trend (or has had the trend removed), we may legitimately regard the sample correlation between the two series as an estimate of the true correlation but we cannot test whether it is significantly different from zero by the ordinary test, since the latter is based on the explicit assumption that successive pairs are statistically independent.

The ordinary statistical models as described above are thus valueless for much serially dependent data and hence we have to construct a new theory in which successive observations are serially dependent on each other. We are therefore led to the theory of random processes. The purpose of this paper is to survey the present state of this theory in so far as it is relevant to the econometrician. Much work not of direct interest in economics is not discussed (e.g. the investigations on spectral analysis by Grenander, Rosenblatt and others) and in addition the choice of subjects is somewhat biased towards those with which I have come in contact, or been instructed in by my colleagues.

## 2. RANDOM PROCESSES WITH ONE VARIABLE

We must first decide whether the random processes we wish to study will involve time in a continuous or a discrete manner. For many applications in physics it is more convenient to take time as continuous and then we have to construct a theory of random functions  $x(t)$ . This can be done, but to obtain a strictly rigorous theory requires a considerable mathematical apparatus. Moreover the data we deal with in economics are always given at discrete intervals. It is true that such data are often not the value of a variate at an instant of time but a sum or integral over an interval of what might be perhaps better regarded as something going on continuously. Nevertheless the simplification resulting from considering discrete moments of time, separated by intervals of constant length, is great and we therefore confine ourselves to this case.

We next have to consider what unit to take for the time interval which might be years, weeks or even days. In most cases a year is the interval considered. This usually results in quite short series, but if we attempt to increase the length of the series by taking months or weeks, we may get into trouble. Not only may we now have seasonal effects to eliminate before further analysis but we find that we are now studying smaller scale phenomena, in the time sense, and if there is a lag or serial dependence between two series, this will be spread over many more intervals than before, usually resulting in much heavier computations. Thus the gain in apparent accuracy by increasing the number of terms may be illusory.



On the other hand, economic effects are probably fairly short term and perhaps of the order of one to three months. One month may therefore be the best unit, if seasonal effects can be removed. In most cases, therefore, we will want to consider a series of variates,  $x_t$ , ( $t = \dots -2, -1, 0, 1, 2, \dots$  say) where the unit in which  $t$  is measured is one year or one month.

If the joint distribution of any set of these  $(x_t, x_u, \dots, x_v)$  depends only on the differences between  $t, u, \dots, v$  and not on their absolute value we say such a process is stationary. In practice the two most usual reasons why this assumption does not apply are the existence of seasonal effects if the time interval is less than one year, and the existence of trend caused by technological or other economic development. The first may be removed by using a correction such as might be obtained from a thirteen month weighted average, or from the totals over the years for the individual months. Neither of these methods can be regarded as very satisfactory but it is probably not possible to find much better methods. A trend may be removed by taking out a linear or higher order regression on time as an auxiliary variable. This introduces the complications connected with regression which will be discussed later. Alternatively we may remove a trend by subtracting a moving average. Besides shortening the series studied this tends to distort the nature of the serial dependence. For the present, therefore, we shall consider processes which are stationary.

For many purposes it is also convenient to assume that the joint distribution of any set of the  $x$ 's is a multivariate normal distribution, but even if this is not assumed we shall always assume that the second moment (and therefore also the first) exists. We may take the first moment or mean to be zero and the second to be  $\sigma^2$ , so that we write  $E(x_t) = 0$ ,  $E(x_t^2) = \sigma^2$ . We then define the serial correlation coefficients,  $\rho_s$ , by

$$\sigma^2 \rho_s = E(x_t x_{t+s}) = \sigma^2 \rho_{-s},$$

and it is convenient to introduce a serial covariance generating function

$$S_x(z) = \sigma^2 \sum_{s=-\infty}^{\infty} \rho_s z^s,$$

where  $z$  is a complex variable (Quenouille (1947), Moran (1949)). This series usually converges in a ring  $1-\delta \leq |z| \leq 1+\delta$  where  $\delta > 0$ , but even if it does not, it can be taken as defining a function on the unit circle  $|z| = 1$ . The advantage of using such a generating function is that if we define a new process  $\{y_t\}$  by an equation of the form

$$y_t = \sum_{i=0}^{\infty} \alpha_i x_{t-i},$$

where  $\sum_{i=0}^{\infty} \alpha_i$  is (say) dominated by a convergent geometric series, then the serial covariance generating function of the  $\{y_t\}$  process is given by

$$S_y(z) = \left( \sum_{i=0}^{\infty} \alpha_i z^{-i} \right) \left( \sum_{j=0}^{\infty} \alpha_j z^j \right) S_x(z).$$

This fact simplifies much of the elementary algebra connected with simple processes.

The fundamental fact about serial correlation coefficients is given by Wold's theorem (1938) which is the analogue for discrete processes of the theorem for continuous processes proved by Khintchine (1934). This states that the necessary and sufficient condition



that any arbitrary sequence of real constants  $\{\rho_k\}$ ,  $k = 0, \pm 1, \dots$  are the serial correlations of some discrete stationary process is that there exist a non-decreasing function  $W(\theta)$  such that

$$W(0) = 0, \quad W(\pi) = \pi,$$

and

$$\rho_k = \frac{1}{\pi} \int_0^\pi \cos k\theta dW(\theta).$$

$\sigma^2 W(\theta)$  is then known as the integrated power spectrum of the process and  $\sigma^2 W'(\theta)$ , if it exists, is known as the spectral density. If  $W'(\theta)$  exists everywhere in the interval  $(0, \pi)$  and satisfies certain very wide regularity conditions, it is given by

$$W'(\theta) = 1 + 2 \sum_{k=1}^{\infty} \rho_k \cos k\theta = S(e^{i\theta}).$$

The idea behind the introduction of the power spectrum comes from the study of processes with continuous time used to represent noise in electrical theory. Here the use of electrical filters leads to the idea that an arbitrary current (possessing some kind of stationarity) can be represented as a sum of periodic components with random amplitudes and phases. The total energy of these currents is the sum of the energies associated with each and the power spectrum is a measure of the proportion of the energy corresponding to each frequency range. When we have a discrete sequence instead of a continuous function, there is an upper bound to the frequencies required and  $W(\theta)$  is only defined for an interval  $(0, \pi)$  whereas in Khintchine's original theorem  $W(\theta)$  has to be defined over the whole interval  $(0, \infty)$ .

Consider now various simple models. The next simplest model after a completely random series is obtained by defining a process  $\{x_t\}$  in terms of a completely random stationary process  $\{\epsilon_t\}$  by the relationship

$$x_t = \rho x_{t-1} + \epsilon_t.$$

This has come to be known as a simple Markov process. It is clear that for  $\{x_t\}$  to be a stationary process we must have  $|\rho| < 1$ , and by multiplying by  $x_{t-s}$  ( $s > 0$ ) and taking expectations we see that  $\rho_s = \rho_{-s} = \rho^s$ . This process possesses what has come to be known as the Markovian character, namely, that if  $x_t$  is known, the conditional distribution of any set of  $x$ 's with suffixes greater than  $t$ , given  $x_t$ , is independent of all the  $x$ 's with suffixes less than  $t$ .

A more general model is obtained by defining  $\{x_t\}$  in terms of  $\{\epsilon_t\}$  by the relationship

$$x_t + a_1 x_{t-1} + \dots + a_k x_{t-k} = \epsilon_t.$$

In order that this generates a stationary process, it is not hard to see that we must impose the condition that all the roots of the equation

$$z^k + a_1 z^{k-1} + \dots + a_k = 0 \quad \dots (2.1)$$

lie inside the unit circle  $|z| = 1$ . The calculation of the serial correlations is a little more complicated but is facilitated by the use of the serial covariance generating function  $S_x(z)$



of  $\{x_t\}$ . Since  $\{\epsilon_t\}$  can be regarded as generated by a moving average of  $\{x_t\}$  and has itself a serial covariance generating function equal to unity, we have

$$(1 + a_1 z + \dots + a_k z^k)(1 + a_1 z^{-1} + \dots + a_k z^{-k}) S_x(z) = 1$$

and so 
$$S_x(z) = (1 + a_1 z + \dots + a_k z^k)^{-1} (1 + a_1 z^{-1} + \dots + a_k z^{-k})^{-1}.$$

The advantage of using this type of model, which was introduced by Yule (1927) and is sometimes known as an autoregressive series, is that by choosing the  $a_s$  so that the roots of (2.1) are complex we can obtain series which show oscillatory behaviour, i.e., a tendency to overswing the mean. In univariate series the models of this type which have been most used have  $k = 2$ .

Another type of model which is easy to discuss is the finite moving average scheme. Here we suppose  $x_t = a_0 \epsilon_t + a_1 \epsilon_{t-1} + \dots + a_k \epsilon_{t-k}$  where  $a_0 > 0$ ,  $a_k > 0$ , and  $\{\epsilon_t\}$  is a completely random series. Then

$$S_x(z) = (a_0 + a_1 z + \dots + a_k z^k)(a_0 + a_1 z^{-1} + \dots + a_k z^{-k}) \dots \quad (2.2)$$

and  $\rho_s = 0$  for  $|s| > k$ . It is sometimes implied in the literature that taking a moving average of a completely random series introduces 'oscillations'. If all the  $a_i \geq 0$ , this is not so, since all the serial correlations are positive if we use the word 'oscillatory' to mean that if the system is disturbed from its mean position there will be a tendency to overswing the mean. Such a genuine 'oscillatory' tendency can only occur, if some of the weights  $a_i$  are negative. What does happen is that a moving average with positive weights will convert an irregular looking series into one which shows an appearance of wandering about the mean and in which neighbouring values tend to be alike. This can be easily mistaken, at first sight, for an oscillatory behaviour and has sometimes led people mistakenly to suppose that apparently oscillatory behaviour can be explained in this way.

Moving averages, especially with equal weights, have sometimes been used to remove trend. Thus given a sequence  $\{x_t\}$  we might estimate a trend value at  $t$  by

$$X_t = (x_{t-k} + x_{t-k+1} + \dots + x_t + \dots + x_{t+k-1} + x_{t+k})(2k+1)^{-1}$$

and then regard  $\{x_t - X_t\}$  as the stationary process to be studied. If  $k$  is not too small, the effect of this on the spectrum is clearly to remove the components with long periods and leave the short-term components relatively unaffected, a linear regression with time being treated as a component of indefinitely long period. Thus the spectrum and therefore the correlational properties of the series is less affected, if we use a moving average of greater extent. Unfortunately economic series are not very long and the use of a long moving average usually means throwing away too many terms at each end.

Finally, another model, which is historically the earliest, is that of 'concealed periodicities.' Here we suppose that  $x_t$  is the sum of a finite number of strictly periodic terms (usually taken as trigonometric) together with a superimposed error, so that we write

$$x_t = \sum_1^k A_s \cos(a_s t + b_s) + \epsilon_t$$

where  $\{\epsilon_t\}$  is a completely random series or may itself be a process with serial dependence of one of the types considered above. This is the natural model for dealing with tidal, meteorological or other geophysical phenomena, in which it is clear that there are strictly periodic



components. It was this idea which led to the introduction and use of the periodogram. Whilst this type of model is very successful in dealing with some phenomena, e.g., the tides, it was its failure with the annual sunspot cycle which led Yule to introduce processes of an oscillatory nature without strictly periodic components and so to introduce random processes into statistics. The processes of concealed periodicities are unlikely to be of much interest to the econometrician except in aiding him in understanding what happens when he removes seasonal variations.

Slutzky (1927) established some interesting theorems on the effects of summing and differencing random series which are often and mistakenly quoted as illustrating a way in which oscillatory behaviour might arise in economic series. If we take a completely random series  $\{\epsilon_t\}$  and take a moving average of two by the formula  $x_t = \epsilon_t + \epsilon_{t-1}$ , and repeat this process a further  $n-1$  times and then take the  $m$ -th difference, we obtain a curve remarkably like a sine wave. In fact if we let  $n$  and  $m$  increase, the series can be represented by a sine wave of period  $L$  given by

$$\cos 2\pi L^{-1} = \frac{n-m}{n+m}$$

to a degree of approximation which gets better and better as  $n$  and  $m$  get larger and larger. By the use of covariance generating functions this can be proved and generalized in a much easier manner than in Slutzky's paper (Moran, 1949 & 1950). This result has in fact no econometric implications, and if we simply take repeated moving averages with positive weights, we do not get an oscillatory process. The heuristic reason for Slutzky's result is as follows. The effect of taking a moving average with positive weights repeatedly is, in effect, to generate a moving average whose weights can be approximated by a multiple of the normal or Gaussian distribution. This is simply a result of the Central Limit Theorem. The effect of taking the  $m$ -th difference is to turn this moving average into one whose weights can be approximated by the  $m$ -th derivative of the normal distribution—the  $m$ -th tetrachoric function. In the main part of its range this function mimics a sine wave (for a proof see Szegő 1939 p. 194). Thus all that the process of adding and differencing has done is to produce a moving average with weights closely graduated by a sine wave. The resulting process thus also mimics a sine wave. It does not appear that this result, however interesting in itself, has any relevance in econometrics.

The main value of the theory of random processes in economics lies in the fact that it enables us to construct models which may be fitted to observed series. However the theory throws some light on economic theory. Samuelson (1947) has shown how economic statics requires a dynamic theory for the discussion of stability problems. Most linear dynamic models, when their parameters are plausibly chosen, produce damped oscillations when they produce oscillations at all. Unless, therefore, we suppose that the systems are essentially non-linear, continuing undamped oscillations of bounded amplitude, such as we require for a trade cycle theory, will have to be explained in some other way. The introduction of a random or stochastic element into a linear model shows how we can set up such an 'endogenous model,' for if we have, say, a simple linear model of the form  $x_t + ax_{t-1} + bx_{t-2} = 0$  with the constants chosen to produce damped oscillations, the process can be kept continually in a state of oscillatory behaviour by disturbing this equation by putting a random element on the right hand side. In this way we can avoid the assumption of non-linear models or



the existence of an exogenous oscillatory factor to account for the trade cycle. This has been known, of course, for a long time. With a more complex model involving two or more variables it is not even necessary to suppose the variables depend on any further variables than those in the immediately preceding time interval, for (as will be seen later) a system of the form

$$x_t = ax_{t-1} + by_{t-1} + \epsilon_t$$

$$y_t = cx_{t-1} + dy_{t-1} + \eta_t$$

can also show oscillatory behaviour.

We now turn to the problem of statistical inference from observed series. We shall consider in turn (1) the estimation of the mean and variance of the process; (2) the testing of serial dependence and the estimation of serial correlations; (3) the estimation of the parameters of autoregressive models and the testing of their goodness of fit; (4) the estimation of the spectrum in general.

Consider first how we might estimate the mean,  $m$  say. Since  $E(x_t) = m$ , we could take the mean of the sample

$$\bar{x} = n^{-1} \sum_{i=1}^n x_i.$$

This has a variance

$$\frac{\sigma^2}{n} \left\{ 1 + 2 \sum_{s=1}^{n-1} \left( 1 - \frac{s}{n} \right) \rho_s \right\} \quad \dots \quad (2.3)$$

which may often be taken in large samples to be

$$\sigma^2 n^{-1} \left\{ 1 + 2 \sum_{s=1}^{\infty} \rho_s \right\}.$$

Here  $\sigma^2$  is  $\text{Var}(x_t)$ . In most economic cases (2.3) will be larger than  $\sigma^2 n^{-1}$ , the value if there is no serial correlation.  $\bar{x}$  is an unbiased estimator but usually not a most efficient one except asymptotically. The most efficient estimator is, in fact, a weighted sum of the  $x_i$ , the weights depending on the serial correlations which have themselves to be estimated. For this reason in most circumstances it is the sample mean which is used. To estimate the variance of the process we may use  $(n-1)^{-1} \sum (x_i - \bar{x})^2$  which is consistent but usually slightly biased. For a further discussion of the estimation of means and variances see Jowett (1955).

We now consider how we can test for serial dependence and the natural thing to do is to use the first order serial correlation coefficient  $r_1$ , which may be defined in a variety of ways giving numerical values which differ only slightly. We might, for example, define the serial correlation coefficient of order  $s$ ,  $r_s$ , to be the product moment correlation between



the pairs of values  $(x_1, x_{s+1}), \dots (x_{n-s}, x_n)$ , but this gives an awkward denominator and it is preferable to use

$$r_s = n(n-s)^{-1} \frac{\sum_{t=1}^{n-s} (x_t - \bar{x})(x_{t+s} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2} \quad \dots (2.4)$$

where  $\bar{x} = n^{-1} \sum_{t=1}^n x_t$ . This is the form used when the true mean of the process is unknown as is almost invariably the case in econometrics. A simpler distributional theory is obtained if we modify this to the 'circular' definition by writing

$$r_s = \frac{\sum_{t=1}^n (x_t - \bar{x})(x_{t+s} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2} \quad \dots (2.5)$$

where we define  $x_{n+t} = x_t$ . A large number of papers have been written on the distributions of the quantities defined by (2.4) and (2.5) both in the null case when the series  $\{x_t\}$  is completely random and under various alternative hypotheses. Good bibliographies will be found in the papers by Watson (1956), Daniels (1956) and Jenkins (1954 and 1956), who also consider partial serial correlations. The exact distributions obtained are invariably awkward analytically, changing their form at a discrete set of points throughout their range, and much effort has gone into providing good approximations. Moreover the effect of correcting for the sample mean in the non-null case is not nearly so simple as for ordinary correlation coefficients.

However a test is not difficult to apply in practice since tables exist for the distribution of  $r_1$  with a suitable definition. For example if, instead of taking (2.4) or (2.5), we define  $r_1$  by

$$r_1 = \frac{\frac{1}{2} (x_1 - \bar{x})^2 + \frac{1}{2} (x_n - \bar{x})^2 + \sum_{i=1}^n (x_i - \bar{x})(x_{i-1} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \dots (2.6)$$

T. W. Anderson (1948) gives one sided 5%, 1% and 0.1% levels of significance for  $n = 4, 5, \dots 60$ . This definition of  $r_1$  can be expressed in terms of von Neumann's mean square successive difference

$$\frac{\delta^2}{S^2} = n(n-1)^{-1} \frac{\sum_{i=1}^n (x_i - x_{i-1})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = 2n(n-1)^{-1}(1-r_1).$$

Tables for the distribution of  $\delta^2/S^2$  were given by B. I. Hart (1942), from which Anderson's were calculated. Alternatively R. L. Anderson (1942) has given tables for the exact distribution of  $r_1$  as given by (2.5). Watson has pointed out in his thesis that the distribution of  $r_1$ , when defined by (2.6) is almost exactly that of the ordinary correlation coefficient as based on  $n+3$  pairs of observations.



T. W. Anderson (1948) has considered in what sense these tests are optimal. If we take the alternative hypothesis that the  $x_t$  are generated by the simple Markov process  $x_t = \rho x_{t-1} + \epsilon_t$  where  $|\rho| < 1$  and the  $\{\epsilon_t\}$  are a completely random series, no uniformly most powerful test exists even for one sided alternatives. However, the test will be close to an optimum one, for all these differing definitions of  $r_1$  give numerical values close to each other, and (2.6) is known (T. W. Anderson (1948) p. 108) to provide a uniformly most powerful one sided test against an alternative hypothesis about the distribution of the  $x$ 's which only differs slightly from the above one. Moreover,  $r_1$  as given by the above definitions is close to the maximum likelihood estimator of  $\rho$  for a simple Markov process generated by  $(x_t - m) = \rho(x_t - m) + \epsilon_t$ , when  $m$  is unknown. Thus tests based on the first order serial correlation coefficient or von Neumann's ratio may be regarded in practice as optimal. However, it should also be noticed that if the series consists of independent random variables with a trend,  $r_1$  will tend to appear significant. To distinguish trend from serial correlation we therefore need carry out a joint regression and serial correlation test which will be discussed later.

It is of some interest to consider the distribution of  $r_1$  in non-null cases. Most of the work done on this has been for a 'circularly correlated' joint distribution of the  $x$ 's. Even approximate forms of this distribution are complicated. However, in the case where  $x_t = \rho x_{t-1} + \epsilon_t$  so that the  $x$ 's have a known zero mean. Jenkins (1954) has given a simple and accurate transformation which gives a very good approximation to the distribution of  $y = \sin^{-1} r_1$ , where

$$r_1 = \frac{\sum_{t=2}^n x_t x_{t-1}}{\sum_{t=1}^n x_t^2}.$$

In this case he shows that the moments are given by

$$\mu'_1 = -\frac{3}{2} \frac{\rho}{(1-\rho^2)^{\frac{1}{2}}} n^{-1} + \frac{1}{8} \frac{\rho}{(1-\rho^2)^{\frac{3}{2}}} (n-2\rho^2)n^{-2} + O(n^{-3}),$$

$$\mu_2 = n^{-1} - \frac{(2-5\rho^2)}{2(1-\rho^2)} n^{-2} + O(n^{-3}),$$

$$\mu_3 = O(n^{-2}), \mu_4 = 3n^{-2} + O(n^{-3}).$$

This appears to be rather more satisfactory than the  $\tanh^{-1}$  transformation introduced by Quenouille (1948, p 262), and is useful when  $|\rho| < 0.9$ . When the true mean is not known these results do not hold even when we make a correction by replacing  $n$  by  $n-1$ . However, if we make this correction, the error is likely to be  $O(n^{-2})$  and little difference may be expected in practice.

It should be noticed that  $r_1$  is, (on definition (2.4) at least) for small samples, quite biased as an estimator of  $\rho$ . This has been shown empirically by Orcutt (1948) and discussed theoretically by Marriott and Pope (1954) and M. G. Kendall (1948). This is a matter of some importance when using an estimated value of  $\rho$  in a significance test for correlation between series, as will be seen later.



A quite different approach to the problem of testing for serial correlation is due to Ogawara (1951). Suppose that the null hypothesis is that the sequence  $\{x_n\}$  is a completely random sequence, normally distributed and the alternative hypothesis is that it is a Markov process, i.e., that it is generated by a relationship of the form  $(x_n - m) = \rho(x_{n-1} - m) + \epsilon_n$  where  $\{\epsilon_n\}$  is a completely random series of normal variates with zero mean, and  $m$  is unknown. This process has the Markov property that the probability distribution of the whole of the series  $\{x_N\}$  for  $N > n$ , given the value of  $x_n$ , is independent of the values of  $x_m$  for  $m < n$ . In fact we can describe it as a 'nearest neighbour' process which means that the conditional distribution of  $x_n$ , given  $x_{n-1}$  and  $x_{n+1}$ , is independent of all the  $x_i$  with  $i < n-1$  or  $i > n+1$ . Ogawara's idea is to calculate, given a series  $x_1, \dots, x_n$  ( $n$  odd), the regression of the  $x$ 's with an even suffix on the means of the neighbouring  $x$ 's. This can then be tested in a  $t$ -distribution in the ordinary way. An equivalent way of doing this test is to calculate the ordinary correlation coefficient between the series  $x_2, x_4, \dots$  and the corresponding values  $(x_1 + x_3), (x_3 + x_5), \dots$ . This has the ordinary distribution of  $r$  as based on  $\frac{1}{2}(n-1)$  pairs of observation. Hannan (1955) showed that this gives a test of the hypothesis  $\rho = 0$  which is asymptotically efficient (for a definition of this expression see Pitman (1948), Noether (1955) and Hannan (1956)), but it is not asymptotically efficient as a test of the hypothesis  $\rho = \rho_0 \neq 0$ . Ogawara's test has the advantage of being exact and having significance levels which are easily found from  $t$ -tables; but since the distribution of  $r_1$  or of von Neumann's ratio has also been tabulated, it does not offer any real advantage in the present case. It can, however, be useful when extended to more complicated situations.

A more important problem is to be able to test for partial serial correlation. If we have a simple Markov process generated by a relationship of the form  $x_n = \rho x_{n-1} + \epsilon_n$  we have  $\rho_s = \rho^s$  and if we define the partial serial correlation coefficient  $\rho_{2.1}$  by the relationship

$$\rho_{2.1} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \quad \dots \quad (2.7)$$

we find  $\rho_{2.1} = 0$ . This is not ordinarily the case for a higher order autoregressive series and so the sample analogue of  $\rho_{2.1}$ ,

$$r_{2.1} = \frac{r_2 - r_1^2}{1 - r_1^2}$$

can be used as a test of the hypothesis that the process is a simple Markov process against the alternative hypothesis that further terms need to be included in the generating relation. This is in fact an approximate likelihood ratio criterion. This method can be extended to higher order schemes to test a null hypothesis of extent  $k$  against one of extent  $k+1$  as originally suggested by Yule. Jenkins (1956) has given a smoothed (i.e. approximate) form for the distribution of such a serial partial correlation coefficient of order  $k$  with the effect of the  $k-1$  intermediate terms removed. The sample serial correlations used are circular and corrected for the mean (if the true mean is known, the results are similar). It is then found that the form of the distribution depends on whether  $k$  is even or odd but the significance level of any observed value can be calculated from tables of the incomplete Beta function. A discussion of the extensive analytical theory required to establish these results will be found in the papers of Watson (1956), Daniels (1956), and Jenkins (1956).



Thus to fit an autoregressive scheme we may calculate in turn successively higher order partial serial correlation coefficients (not forgetting the 'Fallacy of Many Tests') and then having decided on the order, estimate the coefficients of the scheme by least squares using the observed correlations.

It is however possible to go further. Given an estimated autoregressive scheme all the higher order serial correlations can be estimated and if plotted against their order represent the estimated 'correlogram'. The fit of the observed correlogram to this can be tested by goodness of fit criteria, the first of which is due to Quenouille (1947). These tests have been further developed by Bartlett and Diananda (1950), and Walker (1950 and 1952) and extended to two variate processes by Bartlett and Rajalakshman (1953). Such an overall test of goodness of fit<sup>1</sup> may show up inadequacies in a model which do not appear when we simply calculate a few higher order partial correlations. For example Bartlett (1954) has shown that a series of 114 annual values of the logarithms of Canadian lynx trapped in the Mackenzie River district in North West Canada which appear to be well fitted by a second order autoregressive scheme (Moran, 1953) when we consider lower order partial serial correlations, nevertheless show a strongly significant divergence from the estimated correlogram, when goodness of fit test is used. What this implies is not clear.

The failure of models based on concealed strictly periodic elements led to the abandonment of the use of the periodogram in the analysis of processes. However, more recently it has been realised that the periodogram can be a useful method. In fact just as the population correlogram defined by  $\{\rho_s, s = 0, \pm 1, \dots\}$  and the integrated spectrum defined by  $W(\theta)$  are in a kind of Fourier transform relationship to each other given by the Wold-Khintchine theorem, so also are the corresponding sample quantities. If we have an observed series  $\{X_t\}$  with, say, a known zero mean, the periodogram ordinate,  $I_p$ , for a given value of  $p$  is defined by

$$I_p = A_p^2 + B_p^2,$$

where

$$A_p = \sqrt{\left(\frac{2}{n}\right)} \sum_1^n X_t \cos \left( \frac{2\pi pt}{n} \right),$$

$$B_p = \sqrt{\left(\frac{2}{n}\right)} \sum_1^n X_t \sin \left( \frac{2\pi pt}{n} \right).$$

If we take the serial covariance,  $C_s$ , to be equal to  $(n-s)^{-1} \sum_1^{n-s} X_t X_{t+s}$ , we have

$$I_p = 2 \sum_{s=-n+1}^{n-1} \left( 1 - \frac{|s|}{n} \right) C_s \cos \left( \frac{2\pi ps}{n} \right)$$

and

$$E(I_p) = 2\sigma^2 \sum_{s=-n+1}^{n-1} \left( 1 - \frac{|s|}{n} \right) \rho_s \cos \left( \frac{2\pi ps}{n} \right),$$

---

<sup>1</sup>For further developments in hypothesis testing in the analysis of empirical series see Whittle (75), (77).



which shows the relationship between the sample values. Instead of considering the serial correlations we may attempt to estimate the spectrum directly by using  $I_p$ . In most cases the true spectral density exists and is a fairly smooth function. However, the graph of  $I_p$  is usually very irregular, since  $I_p$  and  $I_q$  have approximately zero correlation for  $p, q$  integral. Moreover, as the length of the series increases,  $I_p$  does not converge in probability to its expected values. Various methods of smoothing the periodogram have therefore been suggested by Bartlett (1950 and 1954), Grenander (1951) Grenander and Rosenblatt (1952 and 1954) and others (see also Bartlett and Medhi, (1955)). Having estimated a spectrum we may then consider a goodness of fit test of Kolmogoroff's type to see if it diverges significantly from some estimated spectrum, such as might be obtained, for example, by fitting an autoregressive scheme (Grenander and Rosenblatt, 1954). Unfortunately such tests may be easily upset by the fact that the hypothetical spectrum has to be estimated (Kac, Kiefer and Wolfowitz, 1955).

The above survey of methods of analysing a single variate process is somewhat sketchy but covers rather more than the econometrician is likely to use in practice as he will very rarely have a series of more than 50 terms and often only 20 or 30. He will clearly have to test such series for serial dependence and perhaps estimate a first or even second order autoregressive scheme but with series of such a length the calculation of a periodogram and the testing of the goodness of fit of a periodogram or correlogram would not seem to be very useful. The natural scientist, with long series to deal with and probably also a physical theory of the processes producing such series, will find all the above techniques useful, especially since his prime concern may well be to understand the structure of the process. The econometrician, on the other hand, will be primarily concerned to determine how far his tests of significance and estimation procedures may be upset by the fact that he is not dealing with series of independent variates. Moreover his real concern is usually with the relationships between several series rather than the analysis of a single variate process.

This is perhaps a suitable point to make some remarks about the problem of prediction. Clearly prediction is only possible if serial dependence exists and if we assume that the process is Gaussian, so that if any set of the  $x_t$  is distributed in a multivariate normal distribution, the optimum estimator will be a linear form in all the values of  $x_t$  which have been already observed. This linear form can be estimated by least squares and if the underlying process is an autoregressive one, it will have coefficients equal to the coefficients of the estimated autoregressive relation if the prediction is for one time interval ahead. Prediction for larger intervals ahead can be obtained by repeating the prediction. The errors in such a prediction are of two kinds—the error resulting from not estimating the coefficients in the prediction formula exactly, and the error arising from the fact that the future of the process is not uniquely determined by its past. (Formulae for these errors are given by Stone (1947)). In most cases the latter error is so large that prediction from a single series would be rarely worth while in economics where the serial correlations are usually just large enough to make ordinary statistical methods inapplicable but not large enough to make prediction useful. When predictions may be useful in economics, these are in dealing with processes involving more than one variable. The optimal linear predictor in the sense of least squares is then estimated, as before, by least squares and the same type of theory is applied, only the details of the calculations being more complicated. It should be emphasised that the fact that the



least squares predictor is optimal in the class of all linear predictors does not depend on any assumptions about the underlying structure of the process. It will be optimal in the class of all predictors if the process is a Gaussian one but it is quite possible that some economic processes may be better represented by non-linear relationships and for these non-linear predictors those may be better. The theory of processes with non-linear structures is, however, almost completely unexplored.

A problem of considerable importance to which little attention in this connection has been paid is that of superimposed error. Suppose that an observed process  $\{x_t\}$  is generated by the relations  $x_t = y_t + z_t$ ,  $y_t = \rho y_{t-1} + \epsilon_t$  where  $z_t$  and  $\epsilon_t$  are independent processes of independent variates with zero means and standard deviations  $\sigma_1$  and  $\sigma_2$ . Then  $\{y_t\}$ , which is not observed, is a simple Markov process with zero mean and standard deviation  $\sigma_3$  where  $\sigma_3^2 = \sigma_2^2(1 - \rho^2)^{-1}$ . The process  $x_t$  will have standard deviation  $\sigma_4$  where  $\sigma_4^2 = \sigma_3^2 + \sigma_1^2$  and if we write  $A = \sigma_3 \sigma_4^{-1} < 1$ , its serial correlation coefficients will be  $\rho_s = \rho_{-s} = A\rho^s$  ( $s = \pm 1, \dots$ ). This is not a simple Markov process since, for example,  $\rho_2$  is not equal to  $\rho_1^2$ . The problem of testing whether  $\sigma_1 = 0$  is not an easy one. If one can be sure that the process is of the above form with  $\sigma_1 = 0$  or  $\sigma_1 \geq 0$  one could use the partial serial correlation

$$\rho_{2:1} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$$

as a test criterion. However, this may well become large not because  $\sigma_1 \geq 0$  but because the process is really, say, generated by an autoregressive scheme of higher order. The problem of an adequate test is not therefore solved. However, our troubles do not end here. Suppose we wish to estimate  $\rho$ ,  $\sigma_1$  and  $\sigma_2$ . We now get an identification problem of a peculiar type. For if  $\rho = 0$ ,  $\sigma_1$  and  $\sigma_2$  are individually not identifiable and we can only estimate  $\sigma_1^2 + \sigma_2^2$ . Therefore any attempt to estimate  $\sigma_1$  and  $\sigma_2$  can succeed, only if  $\rho \neq 0$  which we can decide only by a statistical test for serial correlation and this does not give a certain result. Thus whether we should proceed or not with our attempt at estimation depends on a prior test, a situation which raises a rather peculiar problem of statistical inference and may also lead to considerable biases. This situation is made all the worse by the short length of the series with which we are usually concerned. This type of situation must frequently occur in econometrics and deserves much deeper investigation.

The extent of the uncertainty in tests for correlation in small samples does not seem to be widely realized. To illustrate this suppose we have sample series of simple Markov process of lengths 15 and 25 and that the true mean of the processes are known, so that we can use Jenkin's approximate distribution of  $y = \sin^{-1} r_1$  (Jenkins, 1954). We then find that the one sided five per cent points of the distribution of  $\sin^{-1} r_1$  are 0.411 and 0.323 respectively, when  $\rho = 0$ . Let us now consider how large  $\rho$  must be to give a 50% chance of exceeding these values. To do this we find by interpolation the value of  $\rho$  such that  $E(\sin^{-1} r_1)$  is equal to 0.411 and 0.323, and these are 0.440 and 0.337. Thus if we have a series of 15 terms with a serial correlation coefficient less than 0.440, we have more than a 50% chance of not judging the serial correlation significant. But a serial correlation 0.4 can have quite a considerable effect on other tests or methods of estimation we may wish to apply later. This is, of course, nothing other than the dangerous procedure of accepting a hypothesis,



because a test of significance on the data does not reject it. This is very common in econometric work and appears to be unavoidable. An example occurs in Stone (1954). In dealing with a large number of economic series each of 19 terms Stone suggests that the series should be transformed by taking first differences, i.e., if the original series is  $\{x_t\}$ , he uses  $\{y_t\} = \{x_t - x_{t-1}\}$  and suggests, as result of applying von Neumann's ratio, that  $\{y_t\}$  can be approximately regarded as series of independent terms. This of course implies that the original process is not a stationary one. However, since Stone's series are only of 19 terms there might well be a non-negligible serial correlation in  $\{y_t\}$ , which does not show up in the test. Clearly with series of this length the econometrician can only use the best methods available and hopes that the results will not be too far out, but the uncertainty in the procedure should at least cause him to regard his final results with more uncertainty than their nominal standard errors would suggest.

It would seem that many of the methods used in the analysis of economic time series are not very "robust" with respect to serial dependence. This is a more serious matter than their non-robustness with respect to the assumption of normality. If there are serious doubts about the latter something can be done by using a transformation or by the use of parametric methods; for example Wald and Wolfowitz have given a non-parametric test for serial correlation based on the universe of all  $n!$  permutations of the observed values.

Examples in economic literature of serial correlation analysis applied to a single series are not common. Some are given in Davis (1941), Kendall (1946) and Kendall (1953) (stock prices). These are mostly quite long series obtained either by removing a trend with a moving average from annual data as in the case of the Beveridge wheat series, or by using weekly values as in Kendall's second paper. However, the econometrician is rarely faced with analysing a long single series of such a type. His series are usually short and not taken in isolation but in relation to one or more other series. Much of the above theory is therefore not directly useful and has been included here, chiefly because it is a necessary preliminary to the joint analysis of two or more series to which we now turn.

### 3. MANY VARIABLE PROCESSES

We now consider the problem of dealing with situations where at each time instant  $t$  we observe several variables  $x_t, y_t, \dots$  and our approach will be determined by what we decide to regard as variates and what as fixed variables—the distinction made in Section I between correlation problems and regression problems. Thus, to take the simplest case, if we are given a series  $\{x_t, y_t\}$  we may try to analyse this by regarding both  $x_t$  and  $y_t$  as random variables or we may take the values of  $x_t$  (say) as fixed and base our inferences on a model, in which the probability reference set is the joint probability distribution of all the  $y$ 's, the  $x$ 's remaining constant at their fixed observed values. Which of these procedures we adopt depends on the kinds of question we want to answer and also on what we know, from non-statistical evidence, of the nature of the model. Even if it is clear that all the variables are best regarded as variates, it is often useful to base our statistical inferences on the conditional probability distribution obtained by holding some of them fixed.

This distinction is related to another convenient distinction of variables into "endogenous" and "exogenous". This distinction is based on a prior knowledge of the structure of the process. An exogenous variable is one which may influence the endogenous variables but is not influenced by them. In any particular case this distinction is not an absolute



one but is relative to the question asked. Thus  $x_t$ ,  $y_t$  and  $z_t$  might be taken as the temperature, the rainfall and the price of beer. Clearly  $z_t$  cannot affect  $x_t$  and  $y_t$  but may be affected by them. The social scientist will therefore take  $x_t$  and  $y_t$  as exogenous variables and  $z_t$  as endogenous, whilst the physical scientist might also be interested in the relationship between  $x_t$  and  $y_t$ . This distinction is therefore principally important in deciding on the model but also in many cases it is the exogenous variables which are taken to be fixed variables. Whether this is so or not in any prediction problem will depend on whether the prediction involves knowing the future values of the exogenous variables or whether they, also, have to be predicted.

Given, then, a multivariate series of observations we may ask what evidence there is for any dependence between the variables and, if this is forthcoming, what we can say about the kinds of model which may be supposed to have produced the series. Before we can do this, we must consider the theory of mathematical models which could generate the series. We take such models to be either stationary or such that their lack of stationarity comes solely from the variation of the mean with some observed variable or variables—usually time. This variation might be linear or not, but on the removal of its effect the process will be supposed to be strictly stationary. The variates will therefore be taken to have zero means and bounded second moments.

We begin with the case where all the variables are variates and if there are  $p$  of them we represent them by a random column vector  $\mathbf{x}_t$  with  $p$  components and a transpose

$$\mathbf{x}_t' = (x_t^{(1)}, \dots, x_t^{(p)}).$$

For any given  $S (= 0, \pm 1, \dots)$  we define the serial covariance matrix to be

$$(C_s^{jk}) = E\{\mathbf{x}_t \mathbf{x}_{t+s}'\} = (C_{-s}^{kj})' = \begin{pmatrix} C_s^{11} & \dots & C_s^{1p} \\ \vdots & & \vdots \\ C_s^{p1} & \dots & C_s^{pp} \end{pmatrix}$$

where

$$C_s^{jk} = E\{x_t^{(j)} x_{t+s}^{(k)}\} = C_{-s}^{kj}.$$

We then have a theorem due to Cramér (1940) which generalises the Wold-Khintchine theorem and states that the necessary and sufficient conditions that any given set of matrices  $\{C_s^{jk}\}$  ( $s = 0, \pm 1, \dots$ ) be the covariance matrices of a stationary  $p$ -variate process are that there exist  $p^2$  (possibly complex) functions  $W_{jk}(\theta)$  which are defined and of bounded variation for  $-\pi \leq \theta \leq \pi$  and such that

$$C_s^{jk} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-is\theta} dW_{jk}(\theta).$$

For  $j = k$ ,  $W_{jk}(\theta)$  is real and non-decreasing. For all the cases with which we will be concerned the matrix covariance generating function

$$S(z) = \left( \sum_{s=-\infty}^{\infty} C_s^{jk} z^s \right)$$



will be convergent in an annulus  $1-\delta \leq |z| \leq 1+\delta (\delta > 0)$  and  $W'_{jk}(\theta)$  will exist and be given by

$$W'_{jk}(\theta) = \sum_{-\infty}^{\infty} C_s^{jk} e^{is\theta}.$$

Suppose we now define another vector process  $\{y_t\}$  as a moving matrix average of the form

$$y_t = A_0 x_t + A_1 x_{t-1} + \dots$$

where each  $A_i$  is a  $p \times p$  matrix, such that each of the  $p^2$  series formed by the  $(i, j)$  elements of the  $A_i$  is an absolutely convergent series (much wider conditions are sufficient). Then if the matrix covariance generating function of the  $\{y_t\}$  process is denoted by  $S_y(z)$ , we easily see that

$$S_y(z) = \left( \sum_{m=0}^{\infty} A_m z^{-m} \right) S(z) \left( \sum_{n=0}^{\infty} A_n z^n \right).$$

We may define the vector analogue of the stationary autoregressive process in one variate by the equation

$$x_t + A_1 x_{t-1} + \dots + A_k x_{t-k} = \eta_t \quad \dots (3.1)$$

where the  $A$ 's are  $p \times p$  matrices and  $x_t$  is a column vector of variates whose variance-covariance matrix is  $B = (b_{ij})$  and such that the  $\eta_t$ 's for different values of  $t$  are all independent. To ensure that such an equation will generate a stationary process it is necessary to impose some condition on the  $A$ 's and this is found to be that the roots of the equation

$$\left| 1 + \sum_{i=1}^k A_i z^{-i} \right| = 0$$

must all lie outside the unit circle  $|z| = 1$ . When this is satisfied, the matrix covariance generating function of the  $\{x_t\}$  process is found to be

$$\left( 1 + \sum_{i=1}^k A_i z^{-i} \right)^{-1} B \left( 1 + \sum_{i=1}^k A_i z^i \right)^{-1}$$

since this now exists inside some annulus  $1-\delta \leq |z| \leq 1+\delta (\delta > 0)$ .

If we now return to the matrix equation (3.1), we see that by applying suitable operators to the  $p$  scalar equations, which it represents, it is possible to eliminate all the  $x^{(i)}$ 's except one. The resulting equation then looks like an ordinary autoregressive equation on the left hand side (of order at most  $pk$ ) but on the right hand side we will have a moving average of the components of  $\eta_t$ . Thus in general an individual component,  $x^{(i)}$ , in a multivariate system defined by an equation like (3.1) is generated by a process which is not of autoregressive type. Notice also, that it is possible to generate intrinsically oscillatory processes by an equation of type (3.1) when  $k = 1$  which is not the case for a scalar autoregressive model. For further developments in the theory of multiple processes see Whittle (1953).

Mann and Wald (1943) have considered the estimation problem for processes of the above kind. In order to include estimation of the means we write the process in the more general form

$$\mathbf{x}_t + \mathbf{A}_1 \mathbf{x}_{t-1} + \dots + \mathbf{A}_k \mathbf{x}_{t-k} + \mathbf{a} = \boldsymbol{\eta}_t$$

where  $\mathbf{x}_t$  no longer has zero mean. The estimation procedure now depends essentially on the assumption that the  $\boldsymbol{\eta}_t$  are serially independent vectors and in fitting a system of type (3.1) to a set of observations the choice of  $k$  such that the resulting estimated  $\boldsymbol{\eta}_t$  can be regarded as serially independent is a matter of some difficulty which deserves further investigation. However, we suppose  $k$  known and then distinguish between two cases (Mann and Wald). In the first we make the further assumption that the elements of  $\boldsymbol{\eta}_t$  have a diagonal variance-covariance matrix  $(\sigma_{ij})$  which is, of course, unknown. If we were to assume further that the elements of  $\boldsymbol{\eta}_t$  are normally distributed, the maximum likelihood estimators of the  $\mathbf{A}$ 's and  $\mathbf{a}$  are found by minimising the sum of squares  $\sum \boldsymbol{\eta}_t \boldsymbol{\eta}_t'$  and therefore coincide with the least square estimates. However even if the  $\boldsymbol{\eta}_t$  are not normally distributed (but have finite moments), these estimators will be consistent and asymptotically normally distributed with variances and covariances which can be estimated.

Two remarks might be made here which may help the reading of Mann and Wald's paper. The minimisation of the sum of squares  $\sum \boldsymbol{\eta}_t \boldsymbol{\eta}_t'$  is equivalent to minimising the sums of squares of the disturbances in the individual equations. This is no longer true when  $(\sigma_{ij})$  is not a diagonal matrix. Moreover it must be pointed out that the maximisation of the likelihood of the observed series is carried out conditionally. The probability distribution is that of the  $\mathbf{x}_t$  for  $t = 1, \dots, n$  with the values for  $t = 0, -1, \dots, 1-k$  kept fixed. By doing this we avoid the difficulties connected with the Jacobian.

If we assume  $(\sigma_{ij})$  is an arbitrary variance-covariance matrix, we may, assuming an unknown multivariate normal distribution for the components of each  $\boldsymbol{\eta}_t$ , find maximum likelihood estimators in the same kind of way and these are again equal to least squares estimators. Their joint limiting distribution, even if the distribution of  $\boldsymbol{\eta}_t$  is not a multivariate normal, is again a multivariate normal. In both this and the previous case these estimators are asymptotically efficient.

In practice the situation is usually more complicated and we have a system of the form

$$\mathbf{A}_0 \mathbf{x}_t + \mathbf{A}_1 \mathbf{x}_{t-1} + \dots + \mathbf{A}_k \mathbf{x}_{t-k} + \mathbf{B}_0 \mathbf{z}_t + \dots + \mathbf{B}_e \mathbf{z}_{t-e} = \boldsymbol{\eta}_t \quad \dots (3.2)$$

where the  $\mathbf{z}_t$  are vectors of exogenous variates which are regarded as fixed. (This is slightly more general than Mann and Wald's third case). By putting one of the components of  $\mathbf{z}_t$  equal to unity, this includes the case considered above where we have to estimate a mean. We suppose that  $\mathbf{A}_0$  is unknown, but since the process is to be generated sequentially in time by the equation (3.2), we must take  $\mathbf{A}_0$  to be non-singular. Premultiplying by  $\mathbf{A}_0^{-1}$  we get the "reduced" form

$$\mathbf{x}_t + \mathbf{A}_0^{-1} \mathbf{A}_1 \mathbf{x}_{t-1} + \dots + \mathbf{A}_0^{-1} \mathbf{A}_k \mathbf{x}_{t-k} + \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{z}_t + \dots + \mathbf{A}_0^{-1} \mathbf{B}_e \mathbf{z}_{t-e} = \mathbf{A}_0^{-1} \boldsymbol{\eta}_t \quad \dots (3.3)$$

and we can apply the theory of Mann and Wald's second case to estimate  $\mathbf{A}_0^{-1} \mathbf{A}_1, \dots, \mathbf{A}_0^{-1} \mathbf{B}_e$ . The problem now is to return from (3.3) to (3.2).



Clearly this cannot be done without further assumptions. We can write our estimate of equation (3.3) in the form

$$\mathbf{x}_t = \mathbf{C}\omega_t + \epsilon_t \quad \dots (3.4)$$

where  $\omega_t$  is a column vector of all the "predetermined" variables in  $\mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-k}, \mathbf{z}_t, \dots, \mathbf{z}_{t-e}$  arranged in a column,  $\mathbf{C}$  is a  $n \times (k+l)n$  matrix which has been estimated, and  $\epsilon_t = \mathbf{A}_0^{-1} \eta_t$ . From (3.4) we wish to return to

$$\mathbf{A}_0 \mathbf{x}_t = \mathbf{D}\omega_t + \eta_t \quad \dots (3.5)$$

where  $\mathbf{D} = \mathbf{A}_0 \mathbf{C}$ . Consider the  $i$ -th row of  $\mathbf{A}_0$  and call it  $a'_i$ . The corresponding row  $d'_i = a'_i \mathbf{C}$ . If we impose the condition that  $n-1$  of the components of  $d'_i$  are zero ( $n$  is the number of components in  $x_t$ ), we have  $p-1$  equations  $a'_i \mathbf{C}_j = 0$  where  $j$  takes  $n-1$  values and  $\mathbf{C}_j$  is the  $j$ -th column of  $\mathbf{C}$ . Then  $a'_i$  will be determined except for a constant numerical factor. It follows that if in each equation of the original system (3.2) we can prescribe exactly  $n-1$  predetermined variables which do not occur in that equation we have complete identification of  $\mathbf{A}_0$  (except for multiplication by a diagonal matrix) and so we can use least squares methods to solve the problem. This can also be done if the restrictions consist of  $n-1$  more general linear relations.

If there are more than  $n-1$  linear restrictions on some or all of the rows of  $\mathbf{D}$ , the system is "over-identified" and there is no non-zero solution by the above method. It is then necessary to maximise the likelihood of the whole system under these constraints and this leads to much more complicated calculations. If one equation is just identified and some or all of the other equations are over-identified we could proceed as before for the estimation of this single equation but this, which is known as the "limited information" method, is not fully efficient since the extra restrictions on the other equations provide some more information.

A clear account of the issues involved in this problem is given in pp. 292-296 of Stone (1954) and a detailed analysis in Koopmans (1950) (see also Koopmans and Hood (1953)).

The restrictions imposed in order to make equation (3.2) identifiable are usually linear. It may however occur that we have information only on the signs of some of the coefficients. In this case least squares methods would be very difficult to apply and a method of minimisation by using an analogue electronic machine may prove the best method.

Turning away from the general problem of multivariate processes let us consider the simplest problems concerning the relationship between two observed series and ask first how correlation between such series can be tested. The fallacy of ascribing a direct causal connection between variates whose observed correlation is due solely to a common factor is well-known and when this common factor is time, so that both the series of values have a trend, it is unlikely to deceive any statistician. This is in fact the basis of Yule's (1926) first example of a "nonsense correlation" mentioned before. However, even if both series are trend free the ordinary correlation coefficient test of the hypothesis that the correlation is zero cannot usually be applied and owing to the lack of serial independence (in both series) and this fact still seems to be widely unrecognised amongst economists (a typical example of earlier date is Beveridge (1944) p.410). Suppose we have two stationary series  $\{x_n\}$ ,  $\{y_n\}$  with serial correlations  $\rho_s$  and  $\rho'_s$ . Then it is easy to show that asymptotically the variance



of the product moment correlation coefficient between them based on  $n$  pairs of observations is approximately (Bartlett (1935)),

$$(n-1)^{-1} \left\{ 1 + 2 \sum_{s=1}^{n-1} \left( 1 - \frac{s}{n} \right) \rho_s \rho'_s \right\}. \quad \dots (3.6)$$

In most economic applications the second factor in this formula will be substantially larger than unity and the power of the test correspondingly reduced. The exact distribution of  $r$  in the present case is not known; but presumably a good approximation would be to refer it to tables of the ordinary distribution based on

$$1 + (n-1) \left\{ 1 + 2 \sum_{s=1}^{n-1} \left( 1 - \frac{s}{n} \right) \rho_s \rho'_s \right\}^{-1} \quad \dots (3.7)$$

pairs of observations. The difficulty is that we do not know  $\rho_s$  and  $\rho'_s$ . We cannot simply substitute estimates  $r_s, r'_s$  calculated from the observed series, for it can be shown that standard error of sampling of the second factor is comparable to its mean. The best thing to do is to fit an autoregressive series of low order (if this can be done) and calculate the higher order serial correlation coefficients from the coefficients of this relation. If both processes are simple Markovian, we have  $\rho_s = \rho_1^s$  and  $\rho'_s = (\rho'_1)^s$ , so that

$$1 + 2 \sum_{s=1}^{n-1} \left( 1 - \frac{s}{n} \right) \rho_s \rho'_s = n \frac{1 + \rho_1 \rho'_1}{1 - \rho_1 \rho'_1} - 2(1 - \rho_1 \rho'_1)^{-2} \{ (\rho_1 \rho'_1)^n (1 - \rho_1 \rho'_1) - \rho_1 \rho'_1 [1 - n(\rho_1 \rho'_1)^{n-1} + (n-1)(\rho_1 \rho'_1)^n] \}$$

which is asymptotically equal to  $n \frac{1 + \rho_1 \rho'_1}{1 - \rho_1 \rho'_1}$ .

We could then substitute estimates  $r_1$  and  $r'_1$  and proceed as above. The snag about this is that  $r_1$  and  $r'_1$  are usually strongly biased in small samples, as pointed out before, and we may substantially underestimate the variance of  $r$ . However with this precaution the method seems to work fairly well in practice.

A better method is suggested by Quenouille (1949). If we have two Markovian processes  $\{x_n\}, \{y_n\}$ , we may calculate the partial correlation coefficients between  $x_t$  and  $y_t$  with the effects of  $x_{t-1}$  and  $y_{t-1}$  both removed. Hannan (1955) has shown that this procedure results in an asymptotically most efficient test. Quenouille also suggested using a partial correlation between  $x_t$  and  $y_t$  with the effect of only  $x_{t-1}$  (or  $y_{t-1}$ ) removed. Whilst this is a valid procedure, it does not provide an asymptotically most powerful test. Hannan also shows that the use of the crude correlation coefficient,  $r$ , with the "degrees of freedom" corrected by (3.7) is also inefficient.

The alternative hypothesis envisaged here is that the series are generated by equations of the form  $x_n = \rho_1 x_{n-1} + \epsilon_n$ ,  $y_n = \rho_2 y_{n-1} + \eta_n$  with  $\epsilon_n$  and  $\eta_n$  correlated but jointly serially independent for different  $n$ . This test depends essentially on the Markovian character of the series, but if either or both are higher order autoregressive series, a similar procedure could, presumably, be applied. Quenouille checked the validity of his suggestions by a sampling experiment but did not obtain any analytical results about the distribution.



If the alternative hypothesis was that  $y_n = ax_n + \eta_n$  where  $\{x_n\}$  and  $\{\eta_n\}$  are *independent* Markov processes with serial correlations  $\rho_1$  and  $\rho_3$ , Quenouille's test is not fully efficient unless  $\rho_1 = \rho_3$ . It will be noted that  $\{y_n\}$  is not then a Markov process unless this condition is satisfied.

A quite different approach due to Hannan (1955) makes use of the idea of Ogawara's test. Suppose  $\{x_t\}$  and  $\{y_t\}$  are simple autoregressive processes,  $\{y_t\}$  is Markovian and we have a sample of  $2n+1$  successive pairs  $(x_t, y_t)$ . Then we calculate the partial correlation coefficient between  $y_{2t}$  and  $x_{2t}$  ( $t = 1, \dots, n$ ) with the effects of  $(y_{2t-1} + y_{2t+1})$ ,  $x_{2t-1}$  and  $x_{2t+1}$  removed. If the series are normally distributed, this partial correlation coefficient is distributed exactly as an ordinary correlation coefficient based on  $n-3$  pairs of observations. The interest of this test mainly lies in the fact that its distribution is exactly known and does not require a knowledge of the serial correlations of the processes. Hannan (1955) has given a detailed discussion of the asymptotic power of this and the previously considered tests for various alternative hypotheses. From this it is clear that in most cases Quenouille's method is more efficient than Hannan's test, one exception being the case where  $\{x_t\}$  and  $\{y_t\}$  are correlated as a result of their residuals being correlated, and  $\{x_t\}$  is a second order autoregressive process whose first partial serial correlation is large and positive whilst in most cases Hannan's test is more efficient than the use of  $r$  and (3.7).

Consider now what is likely to happen in practice. Suppose we are given two series  $\{x_n\}, \{y_n\}$  which, we assume, are either serially independent or generated by simple Markovian schemes. We wish to test whether the two series are independent or not and therefore have to decide whether both are serially correlated. To illustrate we suppose the true means known and the series to consist of  $n = 15$  or  $25$  terms. Then we have already seen that with  $\rho_1 = 0.440$  ( $n = 15$ ),  $\rho = 0.337$  ( $n = 25$ ) we only have a 50% chance of finding  $r_1$  significant at the 5% level. Suppose we always decide that a series is serially dependent if  $r_1$  does reach the 5% level. Then, if the two series are in fact independent, we have a 75% chance of deciding that at least one of the two series is serially independent and that we can therefore use the ordinary correlation coefficient between the series for which we therefore take as 5% significance level  $\pm 0.497$  ( $n = 15$ ) or  $\pm 0.389$  ( $n = 25$ ). In fact however the true serial correlation is in both series  $\rho_s = \rho_1^s$  and the correct approximate number of degrees of freedom is

$$n \left\{ 1 + 2 \sum_{s=1}^{n-1} \left( 1 - \frac{s}{n} \right) \rho_s \rho_s' \right\}^{-1} - 1$$

(since the true means are known) which for  $n = 15$  and  $25$  turns out to be 9.41 and 19.08 so that our true 5% levels are approximately 0.591 ( $n = 15$ ), 0.432 ( $n = 25$ ). This shows the danger of the procedure here adopted. It would appear that it is better to use Quenouille's method, whenever there is any likelihood at all that the processes may not have been serially independent, provided we can be sure of the order of the autoregression. The efficiency of Quenouille's method in detecting non-independence is almost certainly asymptotically unity and for small samples probably throws away about  $n^{-1}$  of the information.

Finally attention should be drawn to the rather curious fact that in some cases it may be preferable and scientifically plausible to draw deductions about the correlation between series without using any statistical method. Thus to take an extreme case suppose we have re-



corded, by some physical instrument such as an oscillograph, two series which are, as far as the eye can judge, simple sine waves of slightly different frequency. Then one can be sure that they are not directly causally connected or correlated. An actual example similar to this is the question whether the observed cycle in the population of the Canadian lynx (a very definite cycle of about 10 years period) is related to the sunspot cycle (Moran, 1949). A statistical test of this hypothesis, e.g. by using Quenouille's method, would require a good deal of computation. However simple inspection of the data clearly shows that there can be no such dependence, for both cycles are very regular and are sometimes in phase and sometimes out of phase. If the dependence were so strong that the cycle in the lynx was caused by that in the sunspots, this could not happen. On the other hand it is sometimes possible to decide that there is some common cause at work in two series by simple inspection, even though an exact analysis might not even show a significant result. Thus if one compares the production of fox, Canadian lynx and Snowshoe rabbit furs from 1848-1908 given in Brouillette (1934), p. 168, it seems quite clear that some common factor is at work since the main peaks in these series all occur about the same time. Needless to say, such a judgement is best made only by a statistician with experience of the misleading inferences which can be drawn from serially correlated series.

We now turn to the case where regression rather than correlation is the appropriate model. We suppose that  $\{y_t\}$  is the series whose regression on  $\{x_t\}$  is to be examined. If we assume that  $y_t = \alpha + \beta x_t + \epsilon_t$  where  $\{\epsilon_t\}$  is another (unobserved) process, the first thing we want to do is to test whether  $\{\epsilon_t\}$  is serially correlated or not, since our method of analysis in the two cases will be different. The natural thing to do is to calculate the serial correlation coefficient of the residuals and use it as a test criterion. For the case of regression on a single variable series  $\{x_t\}$  the mean and variance depend on the values of the  $x$ 's (Moran, 1950). The exact distribution was investigated by Durbin and Watson (1950 & 1951). They gave two exact bounds for the significance levels the uncertainty being due to the dependence of the distribution on the values of the  $x_t$ . In a later paper Hannan (1957) shows that when the regressor variables are a finite polynomial, the true significance level is very near the upper bound given by Durbin and Watson, (the error being of order  $n^{-2}$  instead of  $n^{-1}$ ). On the other hand if the regressor variables are trigonometric polynomials, an exact solution is known (R. L. Anderson & T. W. Anderson, 1950).

If the serial correlations of the residual process were known, the best linear unbiased estimator for  $\beta$  (and similarly for a multiple regression) could be found by minimising the quadratic form

$$(y - X\beta)' \Gamma^{-1} (y - X\beta)$$

where  $y'$  is the row vector  $(y_1 \dots y_n)$ ,  $\beta$  is the column vector of regression coefficients and  $X$  is the matrix of regressor variables.  $\Gamma$  is the variance-covariance matrix of the residuals. However  $\Gamma$  is usually not known. It is well-known that least squares is not an efficient procedure for estimating the regression coefficients and Watson (1955) and Watson and Hannan (1956) have shown that even a relatively small error in the prescription of  $\Gamma$  can lead to considerable inefficiency in the estimation of  $\beta$  and that in general the use of straightforward least squares ( $\Gamma = 1$ ) is not an efficient procedure unless  $\{\epsilon_t\}$  is serially uncorrelated. However if the regressor variables are smooth functions of time (e.g., polynomials), the least squares estimator is asymptotically efficient. This is just the case where



testing for serial correlation is easiest. The problem of serial correlation in regression analysis has also been treated from the point of view of spectral theory by Grenander (1954).

Another approach to the problem of testing serial correlation in the residuals from a regression is based on Ogawara's idea (Hannan, 1955). If we have two series  $\{y_n\}$  and  $\{x_n\}$  such that  $y_n = \alpha + \beta x_n + \epsilon_n$  and wish to test whether  $\{\epsilon_n\}$  is serially independent against the alternative that it is generated by a simple Markov scheme, we test the partial correlation of the  $\{y_{2t}\}$  with  $\{y_{2t-1} + y_{2t+1}\}$  when the effects of  $\{x_{2t}\}$  and  $\{x_{2t-1} + x_{2t+1}\}$  are removed. This is an asymptotically most efficient test.

This completes what might be reasonably written about those aspects of the theory of stationary random process which are of direct interest to the economist. Needless to say, a great deal of other work on these processes has been published, particularly for processes in which time is taken as a continuous variable, but much of this work is mainly of interest in applications to other sciences.

There are, however, a number of economic problems which involve random processes of special types. Such problems are either concerned with particular economic problems, as for example in Rutherford's (1955) stochastic model to explain income distributions (see also Bernadelli (1944)) or else are concerned with economic planning of particular industrial operations. The latter is a rapidly growing field with a closer relationship to "Operational Research" than to pure economic theory. As examples we may instance the problem of inventory policies (for a survey see Ackoff, (1956)), the programming of hydroelectric systems (Massé (1946)), and the planning of transportation (Beckmann, McGuire and Winsten (1956)).

#### ACKNOWLEDGEMENTS

The above paper could not have been written without the continual advice and criticism of my colleagues, Dr. G. S. Watson and Dr. E. J. Hannan.

#### REFERENCES

- ACKOFF, R. L. (1956): The development of operational research as a science. *Jour. Operational Research Soc. Amer.*, **4**, 265-295.
- ANDERSON, R. L. (1942): Distribution of the serial correlation coefficient. *Ann. Math. Stat.* **13**, 1-13.
- AND ANDERSON, T. W. (1950): The distribution of the circular serial correlation coefficient for residuals from a fitted Fourier series. *Ann. Math. Stat.*, **21**, 59-81.
- ANDERSON, T. W. (1948): On the theory of testing serial correlation. *Skand. Aktuar.* **31**, 88-116.
- BARTLETT, M. S. (1935): Some aspects of the time-correlation problem in regard to tests of significance. *J. Roy. Stat. Soc.*, **98**, 536-543.
- (1950): Periodogram analysis and continuous spectra. *Biometrika*, **37**, 1-16.
- (1954): Problemes de l'analyse spectrale des series temporelles stationnaires. *Publ. de l'Inst. Statist.* (Univ. de Paris), III, 119-134.
- AND DIANANDA, P. H. (1950): Extensions of Quenouille's test for autoregressive schemes. *J. Roy. Stat. Soc. B*, **12**, 108-115.
- AND MEDHI, J. (1955): On the efficiency of procedures for smoothing the periodogram from time series with continuous spectra. *Biometrika*, **42**, 143-150.



# RANDOM PROCESSES IN ECONOMIC THEORY AND ANALYSIS

- AND RAJALAKSHMAN, D. V. (1953): Goodness of fit tests for simultaneous autoregressive series. *J. Roy. Stat. Soc. B*, **15**, 107-124.
- BECKMANN, M., MCGUIRE, C. B. AND WINSTEN, C. B. (1956): Studies in the economics of transportation. Published for the Cowles Commission for research in economics at Yale University, Yale University Press.
- BERKSON, J. (1950): Are there two regressions? *J. Amer. Stat. Ass.*, **45**, 164-180.
- BERNADELLI, H. (1944): The stability of income distributions. *Sankhyā*, **6**, 351-362.
- BEVERIDGE, LORD (1944): *Full Employment in a Free Society*, Allen and Unwin, London.
- BROUILLETTE, B. (1934): *La chasse des Animaux a fourrure au Canada*. Gallimard, Paris.
- CRAMER, H. (1940): On the theory of stationary random processes. *Ann. Math. Stat.*, **41**, 215-230.
- DANIELS, H. E. (1956): The approximate distribution of serial correlation coefficients. *Biometrika*, **43**, 169-185.
- DAVIS, H. T. (1941): *The Analysis of Economic Time Series*. Bloomington Press, Indiana.
- DURBIN J. AND WATSON, G. S. (1950): Testing for serial correlation in least squares regression I. *Biometrika*, **37**, 409-428.
- (1951): Testing for serial correlation in least squares regression II. *Biometrika*, **38**, 159-178.
- GRENANDER, U. (1951): On empirical spectral analysis of stochastic processes. *Ark. Mat.*, **1**, 503-531.
- (1954): On the estimation of regression coefficients in the case of an autocorrelated disturbance. *Ann. Math. Stat.* **25**, 252-272.
- GRENANDER, U. AND ROSENBLATT, M. (1952): On spectral analyses of stationary time series. *Proc. Nat. Acad. Sci., Wash.*, **38**, 519-521.
- (1954): Statistical spectral analysis of time series arising from stationary stochastic processes. *Ann. Math. Stat.* **24**, 537-558.
- HANNAN, E. J. (1955): Exact tests for serial correlation. *Biometrika*, **42**, 133-142.
- (1955): An exact test for correlation between time series. *Biometrika*, **42**, 316-326.
- (1956): The asymptotic powers of certain tests based on multiple correlations. *J. Roy. Stat. Soc. B*, **18**, 227-233.
- (1957): Testing for serial correlation in least squares regression. *Biometrika*, **44**, 57-66.
- HART, B. I. (1942): Significance levels for the ratio of the mean square successive difference to the variance. *Ann. Math. Stat.* **13**, 445-447.
- JENKINS, G. M. (1954): An angular transformation for the serial correlation coefficient. *Biometrika*, **41**, 261-265.
- (1954): Tests of hypotheses in the linear autoregressive model I. *Biometrika*, **41**, 405-419.
- (1956): Tests of hypotheses in the linear autoregressive model II. *Biometrika*, **43**, 186-199.
- JOWETT, G. H. (1955): The comparison of means of sets of observations from sections of independent stochastic series. *J. Roy. Stat. Soc. B*, **17**, 208-227.
- KAC, M., KIEFER, J. AND WOLFOWITZ, J. (1955): On tests of normality and other tests of goodness of fit based on distance methods. *Ann. Math. Stat.*, **26**, 189-211.
- KENDALL, M. G. (1946): *The Advanced Theory of Statistics, II*. Charles Griffin, London, 1946.
- (1948): Note on bias in the estimation of serial correlation. *Biometrika*, **41**, 403-404.
- (1953): The analysis of economic time-series—Part I: Prices. *J. Roy. Stat. Soc. A*, **116**, 11-34.
- KHINTCHINE, A. (1934): Korrelations theorie der stationaren stokastischen prozesse. *Math. Ann.*, **109**, 604-615.
- KOOPMANS, T. C. (1949): Identification problems in economic model construction. *Econometrica*, **17**, 125-144.
- (1950): Statistical inference in dynamic economic models, *Cowles Commission Monograph* **10**, John Wiley, New York.
- AND HOOD, W. C. (1953): Studies in econometric method. *Cowles Commission Monograph*, **14**, 47-49.
- LINDLEY, D. V. (1953): Estimation of a functional relationship. *Biometrika*, **40**, 47-49.
- MANN, H. B. AND WALD, A. (1943): On the statistical treatment of linear stochastic difference equations. *Econometrica*, **11**, 173-220.
- MARRIOTT, F. H. C. AND POPE, J. A. (1954): Bias in the estimation of autocorrelation. *Biometrika*, **41**, 390-402.
- MASSE, P. (1946): *Les Reserves et la Regulation de l'avenir dans la vie Economique*, 2 Vols., Paris, Hermann.



- MORAN, P. A. P. (1949): The statistical analysis of the sunspot and lynx cycles, *J. Animal Ecology*, **18**, 115-116.
- (1949): The spectral theory of discrete stochastic processes, *Biometrika*, **36**, 63-70.
- (1950): The oscillatory behaviour of moving averages. *Proc. Cam. Phil. Soc.*, **46**, 272-280.
- (1950): A test for the serial independence of residuals. *Biometrika*, **37**, 178-181.
- (1953): The statistical analysis of the Canadian lynx cycle, I. *Aus. Jour. Zoology* **1**, 163-173.
- (1956): A significance test for an unidentifiable relation. *J. Roy. Stat. Soc. B*, **18**, 61-64.
- NOETHER, G. E. (1955): On a theorem of Pitman. *Ann. Math. Stat.*, **26**, 64-68.
- OGAWARA, M. (1951): A note on the test of the serial correlation coefficient. *Ann. Math. Stat.*, **22**, 115-118.
- ORCUTT, G. H. (1948): A study of the autoregressive character of the time series used for Tinbergen's model of the economic system of the United States, 1919-1932. *J. Roy. Stat. Soc. B*, **10**, 1-53.
- PITMAN, E. J. G. (1948): *Non-parametric Inference*, Notes for lectures delivered at the University of North Carolina, 1948.
- QUENOUILLE, M. H. (1947): Notes on the calculations of autocorrelations of linear autoregressive schemes. *Biometrika*, **34**, 365-367.
- (1947): A large sample test for the goodness of fit of autoregressive schemes. *J. Roy. Stat. Soc.*, **110**, 123-129.
- (1948): Some results in the testing of serial correlation coefficients. *Biometrika*, **35**, 261-267.
- (1949): Approximate tests of correlation in time series. *J. Roy. Stat. Soc. B*, **11**, 68-84.
- REIERSOL, O. (1945): Confluence analysis by means of instrumental sets of variables. *Ark. Mat. Astr. Fys.*, **32A**, 4, 1-119.
- RUTHERFORD, R. S. G. (1955): Income distributions: a new model, *Econometrica*, **23**, 277-294.
- SAMUELSON, P. A. (1947): *Foundations of Economic Analysis*. Harvard University Press.
- SLUTZKY, E. (1937): The summation of random causes as the source of cyclic processes. *Econometrica*, **5**, 105-146, (translation from a paper published in Russian in 1927).
- STONE, R. (1945): The analysis of market demand. *J. Roy. Stat. Soc.*, **108**, 1-98.
- (1947): Prediction from autoregressive schemes and linear stochastic difference systems. *Proc. Int. Stat. Conf.* **5**.
- (1954): *The Measurement of Consumers Expenditure and Behaviour in the United Kingdom, 1920-1938*. I, Cambridge University Press.
- SZEGŐ, G. (1939): *Orthogonal Polynomials*, New York.
- WALKER, A. M. (1950): Note on a generalisation of the large sample goodness of fit test for linear autoregressive schemes. *J. Roy. Stat. Soc.*, **B**, **12**, 102-107.
- (1952): Some properties of the asymptotic power functions of goodness of fit tests for linear autoregressive schemes. *J. Roy. Stat. Soc. B*, **14**, 117-134.
- WATSON, G. S. (1955): Serial correlation in regression analysis, I. *Biometrika*, **42**, 327-341.
- (1956): On the joint distribution of the circular serial correlation coefficients. *Biometrika*, **43**, 161-168.
- AND HANNAN, E. J. (1956): Serial correlation in regression analysis, II. *Biometrika*, **43**, 436-448.
- WHITTLE, P. (1951): *Hypothesis Testing in Series Analysis*, Almqvist and Wiksell, Uppsala.
- (1952): Some results in Time Series Analysis, *Skand. Aktuar*, **35**, 48-60.
- (1952): Tests of fit in Time Series. *Biometrika*, **39**, 309-318.
- (1953): Estimation and information in stationary time series. *Ark. Mat. Astr. Fys.* **2**, 23.
- (1953): The Analysis of multiple stationary time series. *J. Roy. Stat. Soc. B*, **15**, 125-139.
- WOLD, H. (1938): *A Study in the Analysis of Stationary Time Series*, Almqvist and Wiksell, Uppsala.
- YULE, G. U. (1926): Why do we sometimes get nonsense correlations between time series. *J. Roy. Stat. Soc.*, **89**, 1-64.
- (1927): On a method of investigating periodicities in disturbed series, with special reference to Wolfers's sunspot numbers. *Phil. Trans. Roy. Soc. A*, **226**, 267-298.

Paper received : October, 1957.



# EXPRESSIONS FOR THE LOWER BOUND TO CONFIDENCE COEFFICIENTS

By SAIBAL KUMAR BANERJEE

Indian Statistical Institute, Calcutta

**SUMMARY.** A lower bound to the probability of sample estimate plus (and minus)  $t$ -times ( $t > 1$ ) estimate of sampling error covering the population mean (or total) is derived for samples from non-normal populations. Extensions of the result to the case of ratio estimates and multistage designs are also considered.

## 1. INTRODUCTION

Estimate of sampling variance of estimate indicates variability of the estimate and if the parent population is normal, sample estimate  $x_n$  plus (and minus)  $t$ -times estimate of sampling error covers population mean  $m$  in  $\alpha$  per cent of the cases where  $\alpha$  is defined as

$$\alpha = \frac{1}{\sqrt{n-1}} \cdot \frac{100}{B\left(\frac{n-1}{2}, \frac{1}{2}\right)} \int_{-t}^{+t} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}} dt.$$

For a sample from a non-normal population if  $\hat{m}$  is an estimate of  $m$  and  $\hat{V}(\hat{m})$  (computed from sample readings) an estimate of sampling variance of  $\hat{m}$ , an expression for the lower bound to the probability that  $\hat{m} \pm t\sqrt{\hat{V}(\hat{m})}$  covers  $m$ , may be of some interest. In a paper (Banerjee, 1956) it was shown that if  $x_1, x_2, \dots, x_n$  be a sample of size  $n$  drawn at random with replacement from a population with mean  $m$  and  $B_2$ -coefficient  $B_2$ , then, for  $t > 1$ ,

$$\text{prob. } \left\{ |\bar{x} - m| \leq t \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n(n-1)}} \right\} \geq \frac{1}{\frac{B_2-3}{n} + 1 + \frac{2}{(t^2-1)^2} \left\{ \frac{t^4}{n-1} + 1 \right\}}.$$

An extension of the result to the case of pps sampling in stratified multistage design may be of some interest. Some of the extensions are indicated below which are all based upon a simple lemma. It is seen that the role of  $B_2$  in sampling with equal probability is taken over by a similar parameter in sampling with unequal probability. Stratified design introduces further parameters. The case of ratio estimate (simple ratio for single stratum, and combined ratio for  $k$  strata) is also touched up without assuming bivariate normal distribution

of  $y$  and  $x$ . A 'probability' inequality in  $R \left( \frac{\sum_{i=1}^k M_{iy}}{\sum_{i=1}^k M_{ix}} = \text{population ratio to be estimated} \right)$  of the same form as Fieller's inequality is derived. An extension to multistage design is also indicated.

1.1. Lemma: Let  $\phi(x_1, x_2, \dots, x_p)$  be a function of  $p$  stochastic variates such that  $E\{\phi\} > 0$  and  $E\{\phi^2\}$  exist.



Then

$$\text{prob. } \{ \phi \geq 0 \} \geq \frac{[E\{\phi\}]^2}{E\{\phi^2\}} \quad \dots (1.1.1)$$

1.2. *Proof*: Let a variate  $y$  be defined as

$$\begin{aligned} y &= 1, & \text{if } \phi &\geq 0, \\ &= 0, & \text{if } \phi < 0. \end{aligned}$$

Obviously

$$\phi y \geq \phi,$$

or,

$$E\{\phi y\} \geq E\{\phi\},$$

or,

$$[E\{\phi y\}]^2 \geq [E\{\phi\}]^2, \quad (\because E\{\phi\} > 0)$$

or,

$$E\{\phi^2\} E\{y^2\} \geq [E\{\phi y\}]^2 \geq [E\{\phi\}]^2 \quad (\text{Schwarz's inequality})$$

or

$$\text{prob } \{\phi \geq 0\} = E\{y^2\} \geq \frac{[E\{\phi\}]^2}{E\{\phi^2\}}.$$

## 2. ONE STRATUM PPS SELECTION

2.1. Let there be a finite population consisting of  $N$  units. Let  $y_i$  denote variate value  $y$  of the  $i$ -th unit. Let  $n$  units be selected with replacement from  $N$  units, with probability proportional to some measure of the units. Let  $p_1, p_2, \dots, p_N$  denote the probability for the different units to be selected in a particular draw.

2.2. Let  $z_s$  be an estimate of population total  $M$  as built up from  $s$ -th selected unit. Obviously

$$z_s = \frac{y_i}{p_i}$$

if the  $s$ -th selected unit happens to be the  $i$ -th unit of the population.

2.3. Let us define a function  $L$  of estimators  $z_1, z_2, \dots, z_n$  and  $M$  and  $t^2 (t > 1)$  as

$$L \equiv \frac{t^2 \hat{\lambda}}{n} - (\bar{z} - M)^2 \quad \dots (2.3.1)$$

where

$$\hat{\lambda} = \frac{\sum_1^n (z_s - \bar{z})^2}{n-1}; \text{ and } \bar{z} = \frac{\sum z_s}{n}.$$

2.4. It can be easily shown that

$$\left. \begin{aligned} E\{\hat{\lambda}\} &= \lambda \\ E\{(\bar{z} - M)^2\} &= \frac{\lambda}{n} \\ E\{\hat{\lambda}^2\} &= \frac{B_2 \lambda^2 - 3\lambda^2}{n} + \frac{2\lambda^2}{n-1} + \lambda^2 \\ E[\hat{\lambda}(\bar{z} - M)^2] &= \frac{B_2 \lambda^2 - 3\lambda^2}{n^2} + \frac{\lambda^2}{n} \\ E\{(\bar{z} - M)^4\} &= \frac{B_2 \lambda^2 - 3\lambda^2}{n^3} + \frac{3\lambda^2}{n^2} \end{aligned} \right\} \quad \dots (2.4.1)$$



# EXPRESSIONS FOR THE LOWER BOUND TO CONFIDENCE COEFFICIENTS

where 
$$\lambda = \sum_1^N p_i \left( \frac{y_i}{p_i} - M \right)^2$$

and 
$$B_2 = \frac{\sum_1^N p_i \left( \frac{y_i}{p_i} - M \right)^4}{\lambda^2}.$$

We have accordingly

$$\left. \begin{aligned} E(L) &= (t^2 - 1) \frac{\lambda}{n} \\ E(L^2) &= (t^2 - 1)^2 \frac{\lambda^2}{n^2} \left[ \frac{B_2 - 3}{n} + 1 + \frac{2}{(t^2 - 1)^2} \left\{ \frac{t^4}{n - 1} + 1 \right\} \right] \end{aligned} \right\} \dots (2.4.2)$$

From (1.1.1), (2.3.1) and (2.4.2),

$$\begin{aligned} \text{prob. } \{L \geq 0\} &\geq \frac{1}{\frac{B_2 - 3}{n} + 1 + \frac{2}{(t^2 - 1)^2} \left\{ \frac{t^4}{n - 1} + 1 \right\}} \\ \text{or, prob. } \left\{ \bar{z} + \frac{t\hat{\lambda}}{n} \geq M \geq \bar{z} - \frac{t\hat{\lambda}}{n} \right\} &\geq \frac{1}{\frac{B_2 - 3}{n} + 1 + \frac{2}{(t^2 - 1)^2} \left\{ \frac{t^4}{n - 1} + 1 \right\}} \dots (2.4.3) \end{aligned}$$

2.5. Table 1 below gives numerical values of the lower bound to the probability  $\bar{z} + \frac{t\hat{\lambda}}{n} \geq M \geq \bar{z} - \frac{t\hat{\lambda}}{n}$  for  $t = 3$  and sample size  $n = 4, 6, 8, 10, 12, 20, 30, 50, 100$  for different  $B_2$ -values.

TABLE 1. LOWER BOUND OF PROBABILITY OF THE INEQUALITY  $\bar{z} + \frac{t\hat{\lambda}}{n} \geq M \geq \bar{z} - \frac{t\hat{\lambda}}{n}$   
(values worked out from (2.4.3) taking  $t = 3$ )

(values worked out from (2.1.6) table)									
$B_2$ -value	sample size $n =$								
	4	6	8	10	12	20	30	50	100
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(8)	(10)
1.0	0.727	0.830	0.875	0.899	0.914	0.939	0.951	0.959	0.964
2.0	0.615	0.729	0.789	0.825	0.849	0.897	0.921	0.941	0.955
3.0	0.533	0.650	0.718	0.762	0.793	0.859	0.894	0.923	0.946
4.0	0.471	0.587	0.659	0.708	0.744	0.823	0.868	0.907	0.937
5.0	0.421	0.535	0.609	0.661	0.700	0.791	0.844	0.891	0.929

From Table 1 it is seen that if  $B_2$  be some thing like 4.0 (or less), working with three times the sampling error  $\bar{z} \pm 3\sqrt{\hat{V}(\bar{z})}$  will cover the true value in about 70.8 per cent of the cases (or more) if  $n = 10$ . If, however,  $B_2$  be 3.0 or less (in pps sampling  $B_2$  is likely to be small in general) working again with three times the sampling error  $\bar{z} \pm 3\sqrt{\hat{V}(\bar{z})}$  will cover the true value in about 76.2 per cent of the cases or more.

3. *K* STRATA PPS SELECTION

3.1. Let there be  $K$  finite populations consisting of  $N_1, N_2, \dots, N_k$  units. Let  $y_{ij}$  denote variate value  $y$  of the  $j$ -th unit of the  $i$ -th population. Let us consider a scheme of sampling where  $n_i$  units are selected with replacement with probability proportional to some measure of the units from the  $i$ -th population ( $i = 1, 2, \dots, k$ ). Let  $p_{ij}$  denote the probability that the  $j$ -th unit of the  $i$ -th population will appear in a particular selection while sampling for units from the  $i$ -th population. Obviously,  $\sum_{j=1}^{N_i} p_{ij} = 1$  (for  $i = 1, 2, \dots, k$ ).

3.2. Let  $z_{is}$  be an estimate of population total  $M_i$  of the  $i$ -th population as built up from the  $s$ -th selected unit, among  $n_i$  units selected from the  $i$ -th population. Obviously

$$z_{is} = \frac{y_{ij}}{p_{ij}}$$

if the  $s$ -th selected unit happens to be  $j$ -th unit of the  $i$ -th population.

3.3. Let us define a function  $L$  of estimators  $z_{is}$  ( $s = 1, 2, \dots, n_i$ ;  $i = 1, 2, \dots, k$ ) and  $M_1, M_2, \dots, M_k$  and  $t^2$  ( $t > 1$ ) as

$$L \equiv t^2 \sum_1^k \frac{\hat{\lambda}_i}{n_i} - \left\{ \sum_1^k \bar{z}_i - \sum_1^k M_i \right\}^2 \quad \dots \quad (3.3.1)$$

where

$$\hat{\lambda}_i = \frac{\sum_{s=1}^{n_i} (z_{is} - \bar{z}_i)^2}{n_i - 1}$$

and

$$\bar{z}_i = \frac{\sum_{s=1}^{n_i} z_{is}}{n_i} \quad (i = 1, 2, \dots, k).$$

$$3.4. \text{ We have } L = \sum_1^k \left\{ t^2 \frac{\hat{\lambda}_i}{n_i} - (\bar{z}_i - M_i)^2 \right\} + \sum_{\substack{i,j=1 \\ i \neq j}}^k (\bar{z}_i - M_i)(\bar{z}_j - M_j) \quad \dots \quad (3.4.1)$$

Hence

$$E(L) = (t^2 - 1) \sum_1^k \frac{\lambda_i}{n_i} \quad \dots \quad (3.4.2)$$

where

$$\lambda_i = \sum_{j=1}^{N_i} p_{ij} \left( \frac{y_{ij}}{p_{ij}} - M_i \right)^2 \quad (\text{for } i = 1, 2, \dots, k).$$

3.5. From (3.4.1)

$$\begin{aligned} L^2 = & \left( \sum_1^k l_i \right)^2 + \left\{ \sum_{\substack{i,j=1 \\ i \neq j}}^k (\bar{z}_i - M_i)(\bar{z}_j - M_j) \right\}^2 + \\ & + 2 \left\{ \sum_1^k l_i \right\} \left\{ \sum_{\substack{i,j=1 \\ i \neq j}}^k (\bar{z}_i - M_i)(\bar{z}_j - M_j) \right\} \quad \dots \quad (3.5.1) \end{aligned}$$

where

$$l_i = \frac{t^2 \hat{\lambda}_i}{n_i} - (\bar{z}_i - M_i)^2 \quad (\text{for } i = 1, 2, \dots, k).$$



It can be easily shown

$$E \left\{ \left( \sum_1^k l_i \right)^2 \right\} = E \left[ \sum_1^k l_i^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^k l_i l_j \right]$$

$$E \left\{ \sum_1^k l_i^2 \right\} = (t^2-1)^2 \sum_I^k \left[ \frac{\lambda_i^2}{n_i^2} \left\{ \frac{B_{2i}-3}{n_i} + 1 + \frac{2}{(t^2-1)^2} \left( \frac{t^4}{n_i-1} + 1 \right) \right\} \right]$$

where

$$B_{2i} = \frac{\sum_{j=1}^{N_i} p_{ij} \left( \frac{y_{ij}}{p_{ij}} - M_i \right)^4}{\{\lambda_i\}^2}. \quad \dots (3.5.2)$$

$$E \left\{ \sum_{\substack{i,j=1 \\ i \neq j}}^k l_i l_j \right\} = (t^2-1)^2 \sum_{\substack{i,j=1 \\ i \neq j}}^k \frac{\lambda_i \lambda_j}{n_i n_j} = (t^2-1)^2 \left\{ \left( \sum_1^k \frac{\lambda_i}{n_i} \right)^2 - \sum_1^k \frac{\lambda_i^2}{n_i^2} \right\}. \quad \dots (3.5.3)$$

$$E \left\{ \sum_{\substack{i,j=1 \\ i \neq j}}^k (\bar{z}_i - M_i) (\bar{z}_j - M_j) \right\}^2 = 4 \sum_{\substack{i,j=1 \\ i < j}}^k \frac{\lambda_i \lambda_j}{n_i n_j} = 2 \sum_{\substack{i,j=1 \\ i \neq j}}^k \frac{\lambda_i \lambda_j}{n_i n_j} = 2 \left\{ \left( \sum_1^k \frac{\lambda_i}{n_i} \right)^2 - \sum_1^k \frac{\lambda_i^2}{n_i^2} \right\}. \quad \dots (3.5.4)$$

$$E \left[ \left\{ \sum_1^k l_i \right\} \left\{ \sum_{\substack{i,j=1 \\ i \neq j}}^k (\bar{z}_i - M_i) (\bar{z}_j - M_j) \right\} \right] = 0. \quad \dots (3.5.5)$$

3.6. From (3.5.1)–(3.5.5) it follows

$$\begin{aligned} E(L^2) &= (t^2-1)^2 \left[ \sum_1^k \frac{\lambda_i}{n_i} (B_{2i}-3) + \left( \sum_1^k \frac{\lambda_i}{n_i} \right)^2 + \frac{2}{(t^2-1)^2} \left\{ \left( \sum_1^k \frac{\lambda_i^2}{n_i^2(n_i-1)} \right) t^4 + \left( \sum_1^k \frac{\lambda_i}{n_i} \right)^2 \right\} \right] \\ &= \left( \sum_1^k \frac{\lambda_i}{n_i} \right)^2 (t^2-1)^2 \left[ \frac{\sum_1^k \frac{\lambda_i^2}{n_i^3} (B_{2i}-3)}{\left( \sum_1^k \frac{\lambda_i}{n_i} \right)^2} + 1 + \frac{2}{(t^2-1)^2} \left\{ \frac{t^4 \sum_1^k \frac{\lambda_i^2}{n_i^2(n_i-1)}}{\left( \sum_1^k \frac{\lambda_i}{n_i} \right)^2} + 1 \right\} \right] \end{aligned} \quad \dots (3.6.1)$$

From (1.1.1), (3.3.1), (3.4.2) and (3.6.1),

$$\begin{aligned} \text{prob. } \{L \geq 0\} &\geq \frac{\{E(L)\}^2}{E(L^2)} \\ &\geq \frac{1}{\frac{\sum_1^k \frac{\lambda_i^2}{n_i^2} (B_{2i}-3)}{\left(\sum_1^k \frac{\lambda_i}{n_i}\right)^2} + 1 + \frac{2}{(t^2-1)^2} \left\{ \frac{t^4 \sum_1^k \frac{\lambda_i^2}{n_i^2 (n_i-1)}}{\left(\sum_1^k \frac{\lambda_i}{n_i}\right)^2} + 1 \right\}} \dots \quad (3.6.2) \end{aligned}$$

or,

$$\sum_1^k \bar{z}_i + t \sqrt{\sum_1^k \frac{\hat{\lambda}_i}{n_i}} \geq \sum_1^k M_i \geq \sum_1^k \bar{z}_i - t \sqrt{\sum_1^k \frac{\hat{\lambda}_i}{n_i}}$$

with probability equal to or greater than the right hand most expression of (3.6.2).

3.7. If all the  $n_i$ 's are equal to  $n$  the expression for the lower bound takes the form

$$\frac{1}{n} \frac{\sum_1^k \lambda_i^2 (B_{2i}-3)}{\left(\sum_1^k \lambda_i\right)^2} + 1 + \frac{2}{(t^2-1)^2} \left\{ \frac{t^4}{n-1} \cdot \frac{\sum_1^k \lambda_i^2}{\left(\sum_1^k \lambda_i\right)^2} + 1 \right\} \dots \quad (3.7.1)$$

which is equal to

$$\frac{\theta}{n} \frac{(\bar{B}_2-3)+1}{(t^2-1)^2} + \frac{2}{(t^2-1)^2} \left\{ \frac{\theta t^4}{n-1} + 1 \right\} \dots \quad (3.7.2)$$

where

$$\bar{B}_2 = \frac{\sum_1^k \lambda_i^2}{\sum_1^k \lambda_i} ; \theta = \frac{\sum_1^k \lambda_i^2}{\left(\sum_1^k \lambda_i\right)^2} = \frac{\left\{ \frac{C.V..(\lambda)}{100} \right\}^2 + 1}{K}.$$

Table 2 below gives numerical values of (3.7.2) for  $K = 8$ ;  $n = 4, 8, 12, 16$ ;  $\bar{B}_2 = 2, 3, 4$ , and  $C.V..(\lambda) = 0.0, 25.0, 50.0, 75.0, 100.0, 125.0$ , and  $150.0$ .



# EXPRESSIONS FOR THE LOWER BOUND TO CONFIDENCE COEFFICIENTS

TABLE 2. LOWER BOUND OF THE PROBABILITY

$$\sum_1^k \bar{z}_i + t \sqrt{\sum_1^k \frac{\lambda_i}{n}} \geq \sum_1^k M_i \geq \sum_1^k \bar{z}_i - t \sqrt{\sum_1^k \frac{\lambda_i}{n}}$$

FOR A STRATIFIED DESIGN OF 8 STRATA AND 4, 8, 12 AND 16 UNITS  
PER STRATUM FOR DIFFERENT VALUES OF CV( $\lambda$ )

(values worked out from (3.7.2) taking  $t = 3$ )

$\bar{B}_2$ -value	CV( $\lambda$ ) values as						
	0.0	25.0	50.0	75.0	100.0	125.0	150.0
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
number of units per stratum = 4							
2.0	.905	.901	.890	.872	.848	.819	.786
3.0	.880	.875	.860	.836	.805	.768	.728
4.0	.856	.850	.832	.803	.766	.724	.678
number of units per stratum = 8							
2.0	.943	.941	.936	.928	.917	.903	.887
3.0	.929	.927	.919	.908	.892	.872	.849
4.0	.916	.913	.903	.888	.867	.842	.814
number of units per stratum = 12							
2.0	.953	.952	.949	.943	.936	.927	.917
3.0	.943	.942	.937	.929	.918	.905	.889
4.0	.934	.932	.926	.915	.901	.884	.863
number of units per stratum = 16							
2.0	.957	.957	.954	.951	.945	.939	.931
3.0	.950	.949	.946	.940	.932	.921	.909
4.0	.943	.942	.937	.929	.918	.905	.889

From table 2 it is seen that for a design containing 8 strata and  $n$  units per stratum ( $n \geq 4$ ), con-

fidence statement of the form  $\sum_1^k \bar{z}_i + t \sqrt{\sum_1^k \frac{\lambda_i}{n}} \geq \sum_1^k M_i \geq \sum_1^k \bar{z}_i - t \sqrt{\sum_1^k \frac{\lambda_i}{n}}$  will be true in 76.6 per cent of the cases (or more) if  $\bar{B}_2 \leq 4$  and coefficient of variation of  $\lambda$  values be less than or equal to 100. If the number of strata be increased, other parameters remaining the same, the probability (as judged by the expression for the lower bound) will increase.

3.8. With respect to stratified designs having a constant number of  $n$  units per stratum there is another method of estimation of sampling error and allied confidence interval for the population mean (or total). This method at times may be operationally easy and thus less costly in large scale tabulations. Hence in this context it may not be out of place to discuss that method. Denoting as before by  $z_{is}$  ( $s = 1, 2, \dots, n$ ;  $i = 1, 2, \dots, k$ ) estimate of population total  $M_i$  as built up from the  $s$ -th selected unit from  $n$  units selected from the  $i$ -th population, a set of estimators and a function  $L_1$  may be defined as

$$a_s = \sum_{i=1}^k z_{is} \quad \dots (3.8.1)$$

$$L_1 = \frac{t^2 \sum_{s=1}^n (a_s - \bar{a})^2}{n(n-1)} - (\bar{a} - \sum_1^k M_i)^2 \quad \dots (3.8.2)$$

where

$$\bar{a} = \frac{\sum_{s=1}^n a_s}{n}.$$

It can be easily shown that

$$E(L_1) = (t^2 - 1) \sum_1^k \frac{\lambda_i}{n},$$

and

$$\frac{\{E(L_1)\}^2}{E(L_1^2)} = \frac{1}{\frac{\theta(\bar{B}_2 - 3)}{n} + 1 + \frac{2}{(t^2 - 1)^2} \left\{ \frac{t^4}{n-1} + 1 \right\}} \quad \dots (3.8.3)$$

where  $\lambda_i$ ,  $\bar{B}_2$  and  $\theta$  are as defined earlier in paras 3.4 and 3.7.

Since  $\theta < 1$ , comparing (3.8.3) with (3.7.2) it is seen that (3.7.2) will always be greater than (3.8.3). Hence if judged only by this criterion (viz. the expression for the lower bound of probability of confidence statement being true) confidence statement of the form

$\bar{a} \pm t \sqrt{\frac{\sum (a_s - \bar{a})^2}{n(n-1)}}$  is not to be preferred over  $\sum_1^k \bar{z}_i \pm t \sqrt{\sum_1^k \frac{\hat{\lambda}_i}{n}}$ . If, however, number of units per stratum is large (something like 16 or more) the second method may be used in preference over the first.

#### 4. ONE STRATUM PPS SELECTION RATIO ESTIMATE

4.1. One stratum, pps selection, ratio estimate: Let there be a finite population consisting of  $N$  units. Let  $y_i$ ,  $x_i$  denote respectively variate values of character  $y$  and  $x$  of the  $i$ -th unit. Let  $n$  units be selected with replacement from  $N$  units with probability proportional to some measure of the units. Let  $p_1, p_2, \dots, p_N$  denote the probability for the different units to be selected in a particular draw.

4.2. Let  $z_s$  and  $w_s$  be respectively estimates of population totals  $M_y$  and  $M_x$  of character  $y$  and  $x$  as built up from the  $s$ -th selected unit. Obviously

$$z_s = \frac{y_i}{p_i}; w_s = \frac{x_i}{p_i}$$

if the  $s$ -th selected happens to be the  $i$ -th unit of the population.



# EXPRESSIONS FOR THE LOWER BOUND TO CONFIDENCE COEFFICIENTS

4.3. Let us define a function  $L$  of estimators  $z_1, z_2, \dots, z_n$ ,  $w_1, w_2, \dots, w_n$  and  $M_y$  and  $M_x$  and  $t^2 (t > 1)$  as

$$L \equiv \frac{t^2}{n(n-1)} \left[ \sum_1^n \left\{ z_s - R w_s - (\bar{z} - R \bar{w}) \right\}^2 \right] - (\bar{z} - R \bar{w})^2 \quad \dots \quad (4.3.1)$$

$$= \frac{t^2}{n} \left\{ \hat{\lambda}_z + R^2 \hat{\lambda}_w - 2Rr \sqrt{\hat{\lambda}_z \hat{\lambda}_w} \right\} - (\bar{z} - R \bar{w})^2 \quad \dots \quad (4.3.2)$$

where  $R = \frac{M_y}{M_x}$ ;  $\bar{z} = \frac{\sum_1^n z_s}{n}$ ;  $\bar{w} = \frac{\sum_1^n w_s}{n}$ ;

$$\hat{\lambda}_z = \frac{\sum_1^n (z_s - \bar{z})^2}{n-1}; \quad \hat{\lambda}_w = \frac{\sum_1^n (w_s - \bar{w})^2}{n-1}; \quad r \sqrt{\hat{\lambda}_z \hat{\lambda}_w} = \frac{\sum_1^n z_s w_s - n \bar{z} \bar{w}}{n-1}.$$

4.4. Treating  $z_s - R w_s$  as a variate and taking mathematical expectations of  $L$  and  $L^2$  it can be shown that probability  $\{L \geq 0\} \geq P_0$ ,

where  $P_0 = \frac{1}{\frac{B_2(z, w) - 3}{n} + 1 + \frac{2}{(t^2 - 1)^2} \left\{ \frac{t^4}{n-1} + 1 \right\}} \quad \dots \quad (4.4.1)$

where  $B_2(z, w) = \frac{\sum_1^N p_i \left( \frac{y_i - R x_i}{p_i} \right)^4}{\left\{ \sum_1^N p_i \left( \frac{y_i - R x_i}{p_i} \right)^2 \right\}^2} \quad \dots \quad (4.4.2)$

Hence we have from (1.1.1), (4.3.1) and (4.4.1),

$$\frac{t^2}{n} \left\{ \hat{\lambda}_z + R^2 \hat{\lambda}_w - 2Rr \sqrt{\hat{\lambda}_z \hat{\lambda}_w} \right\} \geq (\bar{z} - R \bar{w})^2 \quad \dots \quad (4.4.3)$$

with probability greater than (or equal to)  $P_0$ .

4.5. Following Fieller (1940) "confidence limits" for  $R$  can be derived from (4.4.3). In brief the method may be indicated as under. From (4.4.3) a quadratic equation in  $R$  and a quadratic inequality in  $R$  can be derived as under :

$$R^2(\bar{w}^2 - p \hat{\lambda}_w) - 2R(\bar{z} \bar{w} - pr \sqrt{\hat{\lambda}_z \hat{\lambda}_w}) + (\bar{z}^2 - p \hat{\lambda}_z) = 0. \quad (4.5.1)$$

$$\text{Quadratic equation : } R^2(\bar{w}^2 - p \hat{\lambda}_w) - 2R(\bar{z} \bar{w} - pr \sqrt{\hat{\lambda}_z \hat{\lambda}_w}) + (\bar{z}^2 - p \hat{\lambda}_z) \leq 0. \quad \dots \quad (4.5.2)$$

$$\text{Quadratic inequality : } R^2(\bar{w}^2 - p \hat{\lambda}_w) - 2R(\bar{z} \bar{w} - pr \sqrt{\hat{\lambda}_z \hat{\lambda}_w}) + (\bar{z}^2 - p \hat{\lambda}_z) \leq 0. \quad \dots \quad (4.5.2)$$

$$\text{where } p = \frac{t^2}{n}.$$

Actual numerical values of  $R$  which satisfy (4.5.1) for clarity of exposition, may be considered under three heads :

$$(a) \quad \bar{w}^2 - p \hat{\lambda}_w > 0; \quad (b) \quad \bar{w}^2 - p \hat{\lambda}_w = 0; \quad (c) \quad \bar{w}^2 - p \hat{\lambda}_w < 0.$$

For (a), roots of equation (4.5.1) are real as the discriminant

$$D = 4\{(\bar{z}\bar{w} - pr\sqrt{\hat{\lambda}_z\hat{\lambda}_w})^2 - (\bar{w}^2 - p\hat{\lambda}_w)(\bar{z}^2 - p\hat{\lambda}_z)\} \\ = 4\bar{z}^2\bar{w}^2\{p^2(C_{ww} - C_{zw})^2 + p(1 - pC_{ww})(C_{ww} + C_{zz} - 2C_{zw})\} > 0$$

where 
$$C_{zz} = \frac{\hat{\lambda}_z}{\bar{z}^2}; C_{ww} = \frac{\hat{\lambda}_w}{\bar{w}^2}; C_{zw} = \frac{r\sqrt{\hat{\lambda}_z\hat{\lambda}_w}}{\bar{z}\bar{w}}.$$

Hence from (4.5.2) limits for  $R$  are

$$R_1 \leq R \leq R_2 \quad \dots (4.5.3)$$

where  $R_1$  and  $R_2$  are the roots of (4.5.1) such that  $R_2 > R_1$ .

For (b), limits for  $R$  are derivable from the relation

$$\bar{z}^2 - p\hat{\lambda}_z \leq 2R(\bar{z}\bar{w} - pr\sqrt{\hat{\lambda}_z\hat{\lambda}_w}). \quad \dots (4.5.4)$$

Under (c) there can arise the sub-cases

$$(c.1) \quad \bar{z}^2 - p\hat{\lambda}_z \geq 0$$

$$(c.2) \quad \bar{z}^2 - p\hat{\lambda}_z < 0.$$

For (c.1) the discriminant of the equation (4.5.1) is positive and as such if  $R_1$  and  $R_2$  be the roots of the equation (4.5.1), limits for  $R$  will be

$$R \leq R_1 \quad \text{or,} \quad R \geq R_2 \quad \dots (4.5.5)$$

where  $R_1$  and  $R_2$  will satisfy the relation

$$R_2 > 0 > R_1. \quad \dots (4.5.6)$$

For (c.2) depending upon the numerical values of  $t$  and  $r$  for given  $\bar{z}^2 - p\hat{\lambda}_z$  and  $\bar{w}^2 - p\hat{\lambda}_w$  the discriminant will be

$$\text{either} \quad (c.21) \quad \text{positive}$$

$$\text{or,} \quad (c.22) \quad \text{negative.}$$

For (c.21) limits for  $R$  will be of the nature (4.5.5). For (c.22) as the roots of (4.5.1) will be imaginary any numerical value of  $R$  will satisfy (4.5.2) and as such limits for  $R$  derivable from (4.4.3) will be  $\infty \geq R \geq -\infty$ .

4.6. Limits of the nature (4.5.5) are practically useless and considering the very nature of the limits, the limits derivable from (4.4.3) cannot strictly be called confidence limits. Such limitations, however, apply equally to the bi-variate approach of Fieller as well.

## 5. $K$ STRATA PPS SELECTION COMBINED RATIO ESTIMATE

5.1.  $K$  strata, pps selection, combined ratio estimate: Let there be  $K$  finite populations consisting of  $N_1, N_2 \dots N_k$  units. Let  $y_{ij}, x_{ij}$  denote respectively variate values of character  $y$  and  $x$  of the  $j$ -th unit of the  $i$ -th population. Let us consider a scheme of sampling where  $n_i$  units are selected with replacement with probability proportional to some measure



# EXPRESSIONS FOR THE LOWER BOUND TO CONFIDENCE COEFFICIENTS

of the units from the  $i$ -th population ( $i = 1, 2, \dots, k$ ). Let  $p_{ij}$  denote the probability that the  $j$ -th unit of the  $i$ -th population will appear in a particular selection while sampling for units from the  $i$ -th population.

Obviously 
$$\sum_{j=1}^{N_i} p_{ij} = 1 \text{ for } (i = 1, 2, \dots, K).$$

5.2. Let  $z_{is}$  and  $w_{is}$  be respectively estimates of population totals  $M_{iy}$  and  $M_{ix}$  of character  $y$  and  $x$  of the  $i$  th population as built up from the  $s$ -th selected unit among  $n_i$  units selected from the  $i$ -th population. Obviously

$$z_{is} = \frac{y_{ij}}{p_{ij}}; w_{is} = \frac{x_{ij}}{p_{ij}}$$

if the  $s$ -th selected unit happens to be the  $j$ -th unit of the  $i$ -th population.

5.3. Let us define a function  $L$  of estimators  $z_{is}$  and  $w_{is}$  ( $s = 1, 2, \dots, n_i$ ;  $i = 1, 2, \dots, k$ ) and  $M_{iy}$  and  $M_{ix}$  ( $i = 1, 2, \dots, k$ ) and  $t^2 (t > 1)$  as

$$L \equiv t^2 \sum_{i=1}^k \frac{\sum_{s=1}^{n_i} (u_{is} - \bar{u}_i)^2}{n_i(n_i-1)} - \left( \sum_{i=1}^k \bar{u}_i \right)^2 \quad \dots (5.3.1)$$

$$= t^2 \sum_{i=1}^k \frac{\sum_{s=1}^{n_i} \left\{ z_{is} - \bar{z}_i - R(w_{is} - \bar{w}_i) \right\}^2}{n_i(n_i-1)} - \left\{ \sum_{i=1}^k \bar{z}_i - R \left( \sum_{i=1}^k \bar{w}_i \right) \right\}^2 \quad \dots (5.3.2)$$

$$= t^2 \sum_{i=1}^k \frac{1}{n_i} \left\{ \hat{\lambda}_{iz} + R^2 \hat{\lambda}_{iw} - 2Rr_i \sqrt{\hat{\lambda}_{iz} \hat{\lambda}_{iw}} \right\} - \left\{ \sum_{i=1}^k \bar{z}_i - R \sum_{i=1}^k \bar{w}_i \right\}^2 \quad \dots (5.3.3)$$

where

$$u_{is} = z_{is} - M_{iy} - R(w_{is} - M_{ix}); \bar{u}_i = \frac{\sum_{s=1}^{n_i} u_{is}}{n_i}$$

$$\bar{z}_i = \frac{\sum_{s=1}^{n_i} z_{is}}{n_i}; \bar{w}_i = \frac{\sum_{s=1}^{n_i} w_{is}}{n_i};$$

$$\hat{\lambda}_{iz} = \frac{\sum_{s=1}^{n_i} (z_{is} - \bar{z}_i)^2}{n_i-1}; \hat{\lambda}_{iw} = \frac{\sum_{s=1}^{n_i} (w_{is} - \bar{w}_i)^2}{n_i-1};$$

$$r_i \sqrt{\hat{\lambda}_{iz} \hat{\lambda}_{iw}} = \frac{\sum_{s=1}^{n_i} z_{is} w_{is} - n_i \bar{z}_i \bar{w}_i}{n_i-1}; R = \frac{\sum_{i=1}^k M_{iy}}{\sum_{i=1}^k M_{ix}}.$$

5.4. It can be easily shown that

$$E(L) = (t^2 - 1) \sum_1^k \frac{\lambda(u_i)}{n_i} \quad \dots (5.4.1)$$

$$\text{and } E(L^2) = (t^2 - 1)^2 \left\{ \sum_1^k \frac{\lambda(u_i)}{n_i} \right\}^2 \left[ \frac{\sum_1^k \frac{\lambda^2(u_i)}{n_i^3} \{B_{2i} - 3\}}{\left( \sum_1^k \frac{\lambda(u_i)^2}{n_i} \right)^2} + 1 + \right. \\ \left. + \frac{2}{(t^2 - 1)^2} \left\{ \frac{t^4 \sum_1^k \frac{\lambda^2(u_i)}{n_i^2(n_i - 1)}}{\left( \sum_1^k \frac{\lambda(u_i)}{n_i} \right)^2} + 1 \right\} \right] \quad \dots (5.4.2)$$

$$\text{where } \lambda(u_i) = \sum_{j=1}^{N_i} p_{ij} \left\{ \frac{y_{ij} - R x_{ij}}{p_{ij}} - (M_{iy} - R M_{ix}) \right\}^2 \quad \dots (5.4.3)$$

$$\text{and } B_{2i} = \frac{\sum_{j=1}^{N_i} p_{ij} \left\{ \frac{y_{ij} - R x_{ij}}{p_{ij}} - (M_{iy} - R M_{ix}) \right\}^4}{\{\lambda(u_i)\}^2} \quad \dots (5.4.4)$$

Hence from (1.1.1), (5.3.3), (5.4.1) and (5.4.2),

$$\text{prob. } \{L \geq 0\} \geq P_0,$$

$$\text{where } P_0 = \frac{\{E(L)\}^2}{E(L^2)} = \frac{1}{\sum_1^k \frac{\lambda^2(u_i)}{n_i^3} (B_{2i} - 3) + 1 + \frac{2}{(t^2 - 1)^2} \left\{ \frac{t^4 \sum_1^k \frac{\lambda^2(u_i)}{n_i^2(n_i - 1)}}{\left( \sum_1^k \frac{\lambda(u_i)}{n_i} \right)^2} + 1 \right\}} \quad \dots (5.4.5)$$

5.5. We have from (5.3.3) and (5.4.5)

$$t^2 \sum_1^k \left\{ \frac{\hat{\lambda}_{iz}}{n_i} + R^2 \frac{\hat{\lambda}_{iw}}{n_i} - 2R \frac{r_i}{n_i} \sqrt{\hat{\lambda}_{iz} \hat{\lambda}_{iw}} \right\} \geq \left\{ \sum_1^k \bar{z}_i - R \sum_1^k \bar{w}_i \right\}^2 \quad \dots (5.5.1)$$

with probability greater than (or equal to)  $P_0$  where  $P_0$  is given by (5.4.5). From (5.5.1) limits for  $R$  can be derived on the same lines as discussed earlier for the case of a single stratum.



# EXPRESSIONS FOR THE LOWER BOUND TO CONFIDENCE COEFFICIENTS

## 6. EXTENSION TO MULTISTAGE DESIGNS

6.1. One stratum, two stage design, pps selection : Let there be a finite population consisting of  $K$  first stage units, where the  $i$ -th unit contains  $N_i$  second stage units. Let  $y_{ij}$  denote value  $y$  of the  $j$ -th second stage unit of the  $i$ -th first stage unit ( $j = 1, 2, \dots, N_i$ ,  $i = 1, 2, \dots, k$ ). Let us consider a two stage sampling scheme where  $n$  first stage units are selected with replacement from  $k$  first units with probability proportional to some measure of the units. Let  $p_1, p_2, \dots, p_k$  denote the probability for the different first stage unit to be selected in a particular draw. Within each selected first stage unit let us select  $n_1$  or  $n_2$  or  $\dots$   $n_k$  second stage units (according as the selected first stage unit happens to be the 1st or 2nd— or  $k$ -th first stage unit) with replacement with probability proportional to some measure of the second stage units. Let  $p_{ij}$  denote the probability that the  $j$ -th second stage unit (of the  $i$ -th first stage unit) having variate value  $y_{ij}$  will appear in a particular selection while sampling for second stage units after the  $i$ -th first stage unit has been selected ( $j = 1, 2, \dots, N_i$ ,  $i = 1, 2, \dots, k$ ). Obviously  $\sum_{j=1}^{N_i} p_{ij} = 1$ , for  $i = 1, 2, \dots, k$

6.2. Let  $z_s$  be an estimate of population total  $M$  as built up from the  $s$ -th selected first stage unit. Obviously,

$$z_s = \frac{1}{p_i} \cdot \frac{1}{n_i} \sum_{t=1}^{n_i} y_{it}(t)$$

if the  $s$ -th selected first stage unit happens to be the  $i$ -th first stage unit and  $y_{i(1)}, y_{i(2)}, \dots, y_{i(n_i)}$  happen to be  $y$ -values of  $n_i$  selected second stage unit within the  $i$ -th first stage unit.

6.3. Let us define a function  $L$  of estimators  $z_1, z_2, z_3, \dots, z_n$ ,  $M$  and  $t^2 (t > 1)$  as

$$L \equiv t^2 \frac{\sum_{s=1}^n (z_s - \bar{z})^2}{n(n-1)} - (\bar{z} - M)^2 \quad \dots \quad (6.3.1)$$

where

$$\bar{z} = \frac{\sum_{s=1}^n z_s}{n}$$

6.4. We have from (1.1.1) and (6.3.1)

$$\begin{aligned} \text{prob.} \quad & \left\{ \bar{z} + t \sqrt{\frac{\sum (z_s - \bar{z})^2}{n(n-1)}} \geq M \geq \bar{z} - t \sqrt{\frac{(z_s - \bar{z})^2}{n(n-1)}} \right\} \\ & \geq \frac{B_2(z) - 3}{n} + 1 + \frac{2}{(t^2 - 1)^2} \left\{ \frac{t^4}{n-1} + 1 \right\} \end{aligned} \quad \dots \quad (6.4.1)$$

where  $B_2(z)$  for the scheme of sampling considered is given as  $\frac{\mu_4(z)}{\{\mu_2(z)\}^2}$ , where

$$\begin{aligned}\mu_2(z) &= \sum p_i \left( \frac{M_i}{p_i} - M \right)^2 + \sum \frac{1}{n_i p_i} \sum p_{ij} \left( \frac{y_{ij}}{p_{ij}} - M_i \right)^2 \\ \mu_4(z) &= \sum p_i \left( \frac{M_i}{p_i} - M \right)^4 + \sum \frac{1}{(n_i p_i)^3} \sum p_{ij} \left( \frac{y_{ij}}{p_{ij}} - M_i \right)^4 + \\ &\quad + 6 \sum \frac{1}{(n_i p_i)^3} (n_i - 1) \left\{ \sum p_{ij} \left( \frac{y_{ij}}{p_{ij}} - M_i \right)^2 \right\}^2 + \\ &\quad + 4 \sum \frac{1}{(n_i p_i)^2} \left( \frac{M_i}{p_i} - M \right) \sum p_{ij} \left( \frac{y_{ij}}{p_{ij}} - M_i \right)^3 + \\ &\quad + 6 \sum \frac{1}{n_i p_i} \left( \frac{M_i}{p_i} - M \right)^2 \sum p_{ij} \left( \frac{y_{ij}}{p_{ij}} - M_i \right)^2.\end{aligned}$$

when  $M_i$  stands for total of  $y$ -values of the  $i$ -th first stage unit i.e.  $M_i = \sum_{j=1}^{N_i} y_{ij}$ , ( $i = 1, 2, \dots, k$ ).

#### REFERENCES

- BANERJEE, S. K. (1956) : A Lower Bound to the Probability of Student's Ratio. *Sankhyā*, **18**, 391-394.  
 COCHRAN, W. G. (1953) : *Sampling Techniques*, John Wiley and Sons, New York.  
 FIELLER, E. C. (1940) : Biological Standardization of Insulin. *Suppl. J. Roy. Stat. Soc.*, **7**, 1-64.  
 ——— (1954) : Symposium on interval estimation. *J. Roy. Stat. Soc.*, **B**, **16**, 175-222.

*Paper received : August, 1957.*



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

By S. J. POTI, M. V. RAMAN, S. BISWAS and B. CHAKRABORTY

*Indian Statistical Institute, Calcutta*

**SUMMARY.** Even in those countries which have shown considerable progress in medical and public health fields morbidity statistics had always remained somewhat inadequate and defective. In India a large mass of the population particularly in the rural areas do not avail of any medical care. It will, therefore, be too much to expect from the data obtained through hospitals and other official agencies to provide adequate and reliable statistics of the morbidity pattern in this country. For obtaining a comprehensive picture of the morbidity conditions one has, therefore, to depend upon other sources particularly sample surveys. The main object of this study was to evolve suitable procedures for the collection of morbidity and medical care statistics by sample surveys.

### CHAPTER 1

#### INTRODUCTION

1.1. The promulgation of the Births, Deaths and Marriages Act in 1886 marked the beginning of registration of vital events on a voluntary basis throughout India. Since then, some of the States have passed Special Acts for the compulsory registration of births and deaths.

1.2. Although three-fourth of a century have elapsed since the enforcement of registration in the country, the machinery is still in its primitive stage with no appreciable improvement and is almost breaking down for apathy and lack of administrative vigilance. The defects inherent in the system still continue without being rectified. Although the reporting of vital events is a primary duty of the people, they are either ignorant of it or indifferent to it. Moreover, there is practically very little use of birth and death certificates at the present time. The chowkidar or the village headman whose responsibility it is to report the vital events occurring in rural areas is already overburdened by his revenue and police assignments with the result that the registration of vital events in rural areas does not get adequate care or attention. This has led to a gross under-registration of births and deaths which may be of the order of 50 per cent. In the urban areas also proper attention has not been paid for improving the system of registration.

1.3. Further, so far as death registration is concerned, the reporting of the cause of death by the chowkidar who is generally illiterate or semi-literate is anything but satisfactory. Apart from a few well-known and easily recognisable communicable diseases like small-pox and plague, the returns obtained are practically of no use for the purpose of health planning or research. The available vital statistics, therefore, are wholly inadequate and unreliable for carrying out any scientific research or effective health planning.

1.4. The deficiency in the existing morbidity statistics is still more glaring. Even in statistically advanced countries, the recording of such events could not be considered as absolutely perfect. Nevertheless, in those countries a ceaseless effort goes on to perfect the machinery responsible for the collection of such statistics by increasing the scope of notifiable diseases to include a larger number of diseases and by supplementing the available



morbidity data by highly specialised surveys. However, a change in the official attitude was discernible in this country since a decade or so and there has been an increasing realisation of the importance and value of vital and health statistics in the fields of health planning and research. The Government of India, realising the magnitude of the health problems, set up a committee in 1943 known as the Health Survey and Development Committee under the chairmanship of Sir Joseph Bhore to review the then existing health conditions and the status of vital statistics in the country, and to suggest ways and means for their improvement. The recommendations contained in the report which was the result of a painstaking and extensive fact-finding study, though accepted in principle, have not, however, been implemented in full. Nevertheless, the Planning Commission of the Government of India has considered some of the aspects implied in the recommendations and introduced them with suitable modifications in the country's development programmes.

1.5. A searching analysis of the available vital statistics has been made by the Committee. In the chapter on 'Vital Statistics', it says, "Preventive and curative work can be organized on a sound basis only on accurate knowledge regarding the diseases and disabilities prevailing in any area. . . . The organization of morbidity statistics for the community presents a difficult problem even in countries in which the development of health services has advanced much more than India, and figures for deaths in view of their greater completeness are generally utilized to a greater extent for the study of health problems, even though the latter constitute more satisfactory material for such study. It is only when adequate medical services covering the whole population and offering protection to all irrespective of their ability to pay for such protection, becomes established and operates over a reasonable period of time, that morbidity statistics of the requisite quality and quantity will develop."

1.6. The available information on morbidity is chiefly confined to those diseases which are made notifiable to the health authorities. This fact together with the inadequate medical care available to the population greatly restrict the extent and accuracy of such returns. The situation has been adequately summed up in the Report of the Health Survey and Development Committee (loc. cit.) which reads, "There are considerable variations in the number of communicable diseases which are notifiable in the different provinces . . . there do not exist, even in the large cities, adequate facilities for ensuring that some of these diseases, for example, tuberculosis, will be notified in sufficient numbers to ensure that a substantial proportion of the actual occurrences will be brought on record."

1.7. The absence of reliable and adequate national statistics of the incidence of diseases and injuries is being keenly felt by the health authorities. Perfection in this direction can be achieved only in the long run. For some years to come one will have to depend on morbidity and mortality statistics collected from 'ad hoc' surveys either on sample basis or a complete investigation in selected areas to ensure reliability and adequacy of the statistical material.

1.8. Number of attempts have been made by public health organizations in recent years to assess the prevalence of certain chronic diseases such as tuberculosis, leprosy etc. Such surveys require trained medical personnel and elaborate laboratory facilities and in a country like India with limited resources the introduction of such investigational methods on a national scale would be beset with practical difficulties. Further, these surveys cover only the so-called chronic diseases and leave entirely the acute diseases which in India form the bulk of the total morbidity.



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

1.9. Apart from the above few studies of morbidity from specific diseases, no comprehensive survey concerning morbidity and medical care on a national basis has hitherto been attempted. The only general health survey ever done in this country was the Singur Health Survey in 1944 which was confined to a small compact rural area in West Bengal. The lack of enthusiasm on the part of the government to initiate comprehensive health surveys may be attributed to the meagre financial resources and to the shortage of trained personnel for carrying out such extensive undertakings. Even if these are forthcoming, a health survey by its very nature will still prove to be a difficult proposition due to the lack of knowledge of a suitable procedure for the collection of health statistics in the peculiar conditions obtaining in the country. Before launching full-fledged morbidity surveys, therefore, it is desirable to initiate small pilot studies and the experience gained by such studies may be utilized with advantage for planning more effectively in the future health surveys. With this end in view, the West Bengal Health Survey was sponsored by the Indian Statistical Institute in 1955, in conjunction with an enquiry into the employment conditions of the rural and urban populations purely as a pilot study to evolve suitable methodology for the collection of morbidity and medical care statistics.

1.10. In recent years, however, the National Sample Survey (NSS), the only organization in India collecting socio-economic statistics on a national scale, has included within the scope of its enquiry a few questions regarding the occurrence of certain vital events like births, deaths, marriages and diseases. The information, though useful for a broad analysis, is not adequate for a critical evaluation of the health conditions of the people. With its unique position in the country, the NSS is, perhaps, the most competent single organization which can undertake health surveys on a nation-wide basis. Advantage should, therefore, be made of this elaborate machinery for conducting health surveys in future for yielding optimum and quick results.

## CHAPTER 2

### SUMMARY FINDINGS

2.1. The West Bengal Health Survey was initiated by the Indian Statistical Institute in 1955 purely as a pilot study to evolve a suitable methodology for the collection of health and medical care statistics. A total of 1172 rural households distributed over 72 sample villages and 566 urban households distributed over 5 sample towns or cities were surveyed. The households selected in the sample were kept under observation for a period of 3 months by periodical visits. The information collected related to

- (i) demographic particulars of the members of the household,
- (ii) housing and sanitary conditions,
- (iii) composition of the diet,
- (iv) morbidity and medical care,
- (v) specific details of current pregnancy terminations and
- (vi) history of past pregnancy terminations of all ever-married women.

2.2. *Morbidity rates.* All illnesses whose duration exceeded 3 months were classified as chronic and those that prevailed for periods shorter than 3 months were



classified as acute. For the former, the morbidity rate has been expressed by its rate of prevalence, that is, the number of cases per 1000 population at an instant of time, and for the latter the morbidity has been expressed by its rate of incidence, that is, the number of new cases arising in a year per 1000 population. The total prevalence rate of chronic diseases observed among the urban population was higher than that observed among the rural population, the rates being 35 and 20 for the urban and rural populations respectively. In respect of acute diseases also the estimated incidence rate of the urban population was higher than that for the rural population, being 423 and 328 respectively.

2.3. Taking all illnesses and injuries together, a total of 100 cases per 1000 rural population and 150 cases per 1000 urban population were observed during the 3-month period of this survey. In a similar survey conducted in U.K. by the Ministry of Health (1946), a total of 5518 cases were observed among a group of 7000 persons during a three-month period, that is, 790 cases per 1000 population. The contrast between the estimates of West Bengal and U.K. is indeed very striking and the results suggest that the people of U.K. are less healthy than those of West Bengal, which contradicts the prevailing notion about the relative levels of health in these two communities.

2.4. Recall lapse could not possibly explain the huge difference between these two reported rates as we have kept a follow-up record for each family over the three-month period by visiting each household 4 times at intervals of 3 weeks. The reason must be largely ascribable to the low level of health consciousness of the West Bengal population which is essentially the result of their low level of living.

2.5. As there is no well-defined line of demarcation between the state of health and that of disease of an individual, it is likely that the morbidity returns obtained in an enquiry of this type will be influenced appreciably by the level of health consciousness of the community. The lower the level of health consciousness, the greater is the chance of overlooking minor ailments and of reporting only those illnesses that cause severe pain or disability.

2.6. In Tables 5.10 and 5.11 are shown the incidence and prevalence rates for each group of diseases. Among acute diseases, those of the respiratory system alone accounted for about half of the morbidity among the rural and urban populations. Dysentery, diarrhoea and other diseases of the digestive system occupy the second place in order of importance in the morbidity pattern of West Bengal.

2.7. Among chronic diseases taken individually, pulmonary tuberculosis, perhaps, is the most important disease particularly in the urban population. The prevalence rates for pulmonary tuberculosis among the urban and rural populations were 3.77 and 1.68 per 1000 persons respectively.

2.8. When the total morbidity for all types of illnesses were estimated, the only important source of error arose due to non-reporting of minor illnesses by the respondents. However, when one attempts to estimate the morbidity rates for individual groups of diseases, the error due to misclassification by causes of disease is bound to arise. The prevailing opinion among public health workers regarding the errors of misclassification is that it is likely to be enhanced considerably if the investigation is conducted by non-medical personnel. As the West Bengal Health Survey was conducted by non-medical investigators, it was considered desirable to assess the validity of the morbidity rates for specific groups of diseases. For this purpose, about 400 addresses of patients attending the O.P.D. of



the R. G. Kar Medical College Hospitals were collected, together with their diagnostic reports. The households to which those patients belonged were apportioned equally between two teams of investigators, one medical and the other non-medical, and the reports obtained by household canvass by these two teams were compared with the corresponding reports obtained from the hospital register. The results of such comparison are shown in Tables 5.1 to 5.7 of this report.

2.9. Two types of discrepancies in tallying the investigators' reports with the corresponding entries in the hospital register were observed, one arising out of failure to report the illness of the afflicted individual and the other arising out of misclassification of the cause of the disease. In respect of the first, the medical investigators' performance was slightly superior to that of the non-medical investigators, the percentage of cases missed being 6.5 and 11.6 respectively. As regards the error due to misclassification also, the overall rate of disagreement for non-medical investigators was higher than that of medical investigators, being 63.1 per cent and 57.3 per cent for the non-medical and medical teams respectively. Hence, from the results of this investigation it is quite apparent that so far as the validity of the disease classification is concerned, reports based on either type of investigating teams are almost equally unreliable. Caution must, therefore, be taken in interpreting the morbidity rates for individual groups of diseases, no matter by whom the investigation is carried out.

2.10. *Disability rates.* Diseases were classified as (i) non-disabling, (ii) disabling but not causing confinement to bed or hospital and (iii) causing confinement to bed or hospital. The second category being not very much recognisable in the case of infants and old age persons who generally have no assigned work, the analysis in these reports relate only to persons in the age group 15-59 years. Out of 332 cases of illnesses observed among the rural population of this age group, 115 cases or 35 per cent were of a non-disabling nature, whereas among the urban population out of 184 cases observed 49 or 27 per cent were non-disabling. Here also, the contrast with the estimates obtained from the health survey conducted in U.K. by the Ministry of Health is very striking. The observed percentage of non-disabling illnesses among the U.K. population was as high as 91 per cent, which again shows how often minor illnesses of a non-disabling nature are not even recognised as an illness by the local population.

2.11. The economic consequences of this can be assessed only if the duration of the disability resulting from such illnesses is taken account of. In the case of the urban population, the number of days lost due to disability from acute and chronic illnesses were 3.36 and 10.10 per year per person respectively, whereas the corresponding rates for the rural population were 3.12 and 4.55 respectively.

2.12. *Medical and Maternity care.* Out of a total of 604 cases of illnesses observed during the three-month period of survey among the rural population only 60 per cent availed of medical care from any recognised system of medicine, whereas out of 351 cases of illnesses observed among the urban population, 68 per cent availed of medical care from one type or another. The corresponding estimate of the percentage of cases which availed of medical care among the population of U.K. as estimated from the data of the health survey carried out by the Ministry of Health was 40 per cent. This does not mean that more often people in West Bengal go in for medical care than in U.K. On the other hand, a substantial number of cases of illnesses remains unreported due to the failure of the afflicted individuals to recognise the disease.



2.13. Of the three important systems of medicine practised in this country, namely, Allopath, Homeopath and Ayurved or Unani (Indian), the one most frequently resorted to is the allopathic system, 41 per cent of the cases among the rural and 51 per cent of the cases among the urban population availing of this system.

2.14. Those who did not avail of any sort of medical treatment were asked to state reasons for not availing medical care. 41 per cent of rural and urban patients stated that 'sickness was not serious' and 33 per cent of the rural and 6 per cent of the urban patients stated that 'medical care was too expensive'.

2.15. As regards maternity care, the present position seems to be very unsatisfactory. About 24 per cent of the rural deliveries and 20 per cent of the urban deliveries were without any sort of professional attendance. Those attended by untrained nurses (dhais) comprised 76 per cent of the rural deliveries and 25 per cent of the urban deliveries.

2.16. *Infant mortality.* For the purpose of relating the general health of the population to various factors such as nutrition, housing and sanitary condition etc., it was thought advisable to use one single index such as infant mortality rate. An analysis of infant mortality rate by level of nutrition and sanitation of the dwelling place shows a high association of these factors with the infant mortality rate. It was observed that among the class of population in a moderately good level of nutrition, the estimated infant mortality rate was 116 for the rural and 111 for the urban population, whereas among the under-nourished class, the estimated infant mortality rates were 171 for the rural and 150 for the urban population. Similarly, it was observed that among those in the urban sector whose housing condition was somewhat satisfactory the infant mortality rate was 85 as against 149 recorded among those whose housing condition was definitely unsatisfactory. Infant mortality rates were also analysed by certain socio-economic characteristics of the fathers of the infants and here also it was observed that the association of infant mortality rates to these factors was very striking. Though these factors may not directly determine the health conditions of the population, they serve as useful criteria in stratifying the population for obtaining optimum results in a general health survey of this kind.

2.17. *Reliability of the estimates.* It should be emphasised at the outset, that this survey being in the nature of an exploratory one, the estimates are not claimed to be very reliable. However, it would be useful to have an approximate assessment of the degree of accuracy of the estimates, at least of the marginal ones, as such estimates of error will help to formulate suitable sample designs in future. As this survey was conducted in conjunction with an employment survey no modification of the NSS design to suit specifically the collection of health statistics was attempted. It can, therefore, be expected that if in future the sample design is made in such a way as to conform to the requirements of an investigation of this kind, a greater degree of accuracy can be attained even with the same sample size.

2.18. The entire rural and urban samples in this survey were divided into two sets of interpenetrating sub-samples and some of the important estimates given in the body of the report have been obtained sub-samplewise and shown in Appendix 1.



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

2.19. It will appear from these estimates that the rural sub-samples which were comparatively large in size showed fairly close agreement in at least the marginal estimates. The urban sub-samples, however, showed considerable disparity even in the marginal estimates, which may be partly due to the smallness of the sample size and partly due to the greater heterogeneity among the urban sample units. The latter point clearly indicates that a more effective stratification is called for in the urban areas to obtain optimum results.

### CHAPTER 3

#### DESIGN OF THE SURVEY

3.1. As the present survey was undertaken with the main objective of developing a suitable methodology for the collection of health and medical care statistics, its coverage and scope were necessarily limited to that of a pilot study. This pilot study covered the rural and urban populations of West Bengal. For the sake of administrative convenience, the design adopted in the fourth round of the NSS was adhered to. The rural sample consisted of 1172 households spread over 72 villages selected from 16 strata and the urban sample consisted of 566 households spread over 40 town or city blocks in 4 towns and Calcutta City selected from 3 urban strata of which Calcutta City alone constituted one stratum. The rural and urban samples were then split into 2 interpenetrating sub-samples and allotted to separate teams of investigators.

3.2. The State of West Bengal has a population of about 26 million persons of which three-fourths live in rural areas and one-fourth in urban areas. The density of population is approximately 800 persons per square mile. The number of villages exceeds 35,000 and there are over 100 towns with population below 1,00,000 and 7 cities with population over 1,00,000. The percentage of literacy is of the order of 20. In addition to the above demographic features which characterise most of the densely inhabited areas of India, the large influx of refugees in recent times has accentuated the socio-economic problems of the State. Further, being one of the most industrialized States in the Indian Union, the health problems confronting a community as a result of industrialization are a special feature of this population. The study of the various aspects of this community therefore, may be fruitful for application on a wider scale in future health surveys.

3.3. While the estimates obtained in the course of the analysis of the data can be taken as indicative of the general pattern or trend of certain vital events, no claims are put forward as regards the reliability of such estimates.

3.4. The items of information collected in this survey pertain to

- (i) Demographic particulars of the members of the household,
- (ii) housing and sanitary condition,
- (iii) composition of diet,
- (iv) morbidity and medical care,
- (v) specific details regarding pregnancy terminations taking place during the reference period and
- (vi) history of past pregnancy terminations of married women of the household.



3.5. The exact details of information falling under the above 6 main heads can be seen from the schedule, a facsimile of which is reproduced in Appendix 2.

3.6. *Method of investigation.* In general, there are two lines of approach to the study of illness in a population.

- (i) The single visit survey by which records of illness for a sample population on the day of visit or for a specified period of time previous to the visit are collected and
- (ii) keeping a sample population under continued observation over a period of time and recording of illnesses happening during that time.

3.7. Though both the methods yield valuable results, it has to be borne in mind that due to limitations of memory of the informants some of the illnesses particularly those of the respiratory and digestive systems which are of a minor nature and cause little or no disability are largely missed unless the period referred to is a short one. In any survey where probing into past events is a necessary feature of the investigational procedure, due safeguards have to be taken to eliminate or reduce to the minimum the effects due to memory lapse. This is an intricate problem as too short a period will considerably reduce the time coverage and too long a period will obviously result in serious recall lapse. Even in countries where the public health administration has attained a high level of efficiency the element of memory lapse has added considerably to the difficulties of carrying out morbidity surveys. The investigational procedure of this survey was designed in such a fashion that it was possible for the investigators to visit each household selected in this survey 4 times during the period of the investigation. The duration of observation extended over a period of 3 months commencing 14 days prior to the first visit and terminating with the date of last visit. As the period between successive visits was usually 3 to 4 weeks, a continuous record of vital events could be collected by this survey which may be reasonably assumed to be free from any major source of error arising out of recall lapse and at the same time providing a substantial coverage over time.

3.8. The schedule has been divided into 9 blocks and each block relates to a specific aspect of the survey. In the first visit all the blocks except block 9 were to be filled in. In the second and third visits only the information regarding the illnesses or injuries and medical care (block 7) and current terminations (block 8) were to be entered. In the fourth visit which was the final one, besides blocks 7 and 8, block 9 giving particulars of past terminations was to be filled in.

3.9. To avoid possible errors arising out of vague definitions of terms it is essential at the outset of any survey, to give precise definitions to certain categories included in the questionnaire. It may be even desirable at times to modify the conventional definitions of terms to suit the specific object in view, or to conform to the investigational procedure. For instance, the usual definition of a household as denoting a group of persons taking principal meals from a common kitchen for at least 16 days during the month preceding the date of enquiry, which is generally applied to socio-economic studies had to be slightly extended for the purpose of this survey to include within its scope three more categories, namely, (i) all children born during 14 days prior to the date of visit to members of the household, (ii) all persons dead during the 14 days prior to the date of visit, who, if alive, would have been classed as members of the household and (iii) all persons admitted into



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

hospitals and other institutions who, otherwise, would have been classed as members of the household.

3.10. The demographic characteristics of the household and its members are included in blocks 2 and 6 of the schedule respectively. It may be seen that block 6, besides containing items relevant to the present health study, contains additional demographic particulars regarding the individual members. These were collected in connection with the employment survey about which a reference has been made earlier. Only items falling under columns 1 to 5, 7, 14, 15 and 18 to 20 have been considered for the purpose of this study as these alone have a direct bearing on the health aspects.

3.11. The belief that some knowledge of the behaviour of certain factors like the weight, temperature, fatigue and appetite of individuals might eventually give a clue to hidden cases of tuberculosis prompted us to elicit information on these points and they were included in columns 18 to 20.

3.12. Carrying out a regular diet survey is beset with practical difficulties. In the absence of such a survey, the best approach to assess the nutritional level of the population would be to evolve an index which would be indicative of such a level. This was made possible by ascertaining the composition of diet consumed by the household in broad categories during the month previous to the first visit. Block 4 gives particulars in this respect.

3.13. Housing is as important as nutrition. Men spend, on an average, about one half of their time at home, women and children more. In block 5 were to be entered particulars of housing and environmental conditions.

3.14. Block 7 is meant for recording information regarding morbidity, mortality and medical care. These data were to be collected and recorded in all the 4 visits to the household. For each case of illness or injury a separate line was to be used in each visit. If more than one illness occurred to the same person at different times during the reference period, they were to be considered as different cases and entered in different lines. If, however, more than one illness operated simultaneously, then the entries were to be made in a single line. Any illness which prevailed during the time of visit was to be marked as 'P' in column 7 and followed up by entries in a separate line in the next visit. In each visit, new cases occurring in the respective reference periods were to be entered as 'new' and cases carried forward from previous visits were to be entered as 'old' in column 2.

3.15. For each new case, the informant was asked to state the name of the disease as known to him and this was entered exactly in the same manner as given by him. When such names were given in the informant's own language, the investigator was not expected to substitute them by their English equivalents as it was thought that such a procedure might lead to misclassification of diseases, when non-medical investigators were employed for this purpose. Also the investigators were instructed not to ask questions regarding signs and symptoms of the disease, because it was feared that interrogations by non-medical investigators were bound to be incoherent and hence confusing to the respondent. If, however, details about signs, symptoms etc., of the disease were given by the informant of his own accord, they were then to be recorded in column 15, meant for the 'remarks'.

3.16. Diseases have been broadly divided into two classes, namely, acute diseases (duration not exceeding 3 months) and chronic diseases (duration exceeding 3 months). Such diseases which started during the reference period and continued throughout were



classified as acute or chronic with the help of expert medical advice. Each of these two classes was further subdivided into 3 groups according to the nature of disability caused, namely, (i) non-disabling, (ii) disabling but not causing confinement to bed or hospital and (iii) causing confinement to bed or hospital. In all non-bed cases disability to perform normal function as working, attending school etc., due to sickness was a readily recognizable fact, wherever it existed.

3.17. However, in cases of illness to those persons who have no such functions to perform as for instance, infants and aged persons, it is rather difficult to distinguish between the state of disability and non-disability. In such cases, the sickness was considered as disabling if the affected persons availed of either medical treatment or special diet. Here also, as before, the diseases were classified into 6 categories according to the nature of disability caused and the appropriate codes were entered in column 5.

3.18. The date of onset of a disabling disease was reckoned from the date on which the disability actually started and was entered in column 6. If the disease was a non-disabling one, no information on date of onset was sought. The date of recovery was the date on which disability ceased, and if the patient recovered within the reference period of a particular visit, the date of recovery with the prefix 'R' (for example, R—4th April) was entered in this column. If the illness resulted in death, the date of death with the prefix 'D' (for example, D—4th April) was entered in this column. If illness was prevailing on the date of visit merely the letter 'P' was entered in this column, indicating that a follow-up was needed in the next visit.

3.19. The type of medical attendance availed for the illness under consideration fell under the categories of 'allopath', 'homeopath', 'ayurved', 'unani' and 'quack'. Appropriate codes were entered in column 9. In view of the fact that a significant proportion of the population, particularly in the lower social and economic levels, still seek or are forced by circumstances to seek the help of quacks, it was thought desirable to add this category also to the various types of medical attendance, though in any scientific discussion such cases are to be taken as equivalent to no medical care availed. In those cases where more than one type of medical help was availed, multiple codes were to be entered and those cases which were not medically attended were to be shown as having not received medical attendance. Obviously, for those belonging to the latter category, the attendance would not arise.

3.20. Details of expenditure on medical care incurred for each case during each reference period were to be entered in columns 10 to 12. Physician's fees and cost of medicines were to be entered in columns 10 and 11 respectively while column 12 was meant for writing down such expenses incurred towards hospital rent, nursing, transport etc. If the amount expended by the household was to cover more than one case of illness, then it was necessary to split the total amount and to allocate to each case its share.

3.21. It is generally known that a sizeable fraction of illnesses does not receive any medical attention at all. Such being the case, it was thought useful to elicit information as to why medical care was not sought. The probable answers such as 'hospital or physician not available', 'too expensive', 'no faith in treatment', 'sickness not serious', were codified and the appropriate code (s) was entered in column 13. If there were reasons other than those specified above, they were all lumped together and put under the general head 'other reasons' and given a separate code.



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

3.22. Name of the disease as stated by the informant and entered in column 4 together with the available details regarding the signs and symptoms of the disease given by the informant of his or her own accord formed the principal criteria for classification of diseases by causes. It was felt at the outset of the survey that without sufficient confirmatory evidence from factual experience it would be difficult to assess the accuracy or validity of the information on diseases thus collected. Hence, an attempt was made in this investigation itself to collect relevant material for a validity study. For this purpose, investigators were instructed to pick out cases which were medically attended and to enter the diagnostic report of the attending physician wherever such reports were accessible. Such a check was expected to give the necessary supporting evidence for the validity of the returns obtained in the survey.

3.23. Only such cases of child-birth occurring to members of the selected households during the four reference periods were to be entered in block 8. As the number of households covered by the sample was of the order 1750, it was not expected that more than 80 births would be recorded during the course of the observational period of 3 months. Hence, information on only a limited number of items pertaining to post-natal care have been collected.

3.24. The serial number of the women as given in block 6, whose pregnancy terminated during the reference period was entered in column 3 of this block and the nature of termination (live birth, still birth or abortion) and the date of termination were recorded in columns 4 and 5 respectively. If the termination resulted in live birth, the sex of the child was entered in column 6 and information regarding its survival and presence in the household at the time of visit was entered in column 7. The age of the child at the time of visit, or at death or at departure from the household, whichever was applicable, was recorded in column 8.

3.25. Since there was very little time-lag between the visit and the actual happening of the event, it was possible to obtain information on some more aspects pertaining to the event to a greater degree of accuracy than what it would have been if the event related to the remote past. The place of delivery, such as, this household, another household, hospital etc., and the type of attendance, such as, doctor, midwife or nurse, *dhai*, hospital, relative or friend, were to be recorded in columns 9 and 10. The definition of confinement adopted in this survey was the period of hospitalisation or in cases of home delivery, the period of bed-days and the period following confinement and terminating with the resumption of normal duties of the woman was termed as convalescence. These were entered in columns 11 and 12 respectively. The expenditure incurred towards confinement was to be distributed over 'physician's fees', 'midwife or *dhai*'s fees', 'cost of medicines, tonics, etc.' and 'hospital charges'. These were entered in columns 13 to 16. In case the termination resulted in the death of mother, this was to be noted down in column 18. Any other relevant information pertaining to child-birth was to be put down by the investigator in the column headed 'general remarks'. If at the time of any visit the period of confinement or convalescence had not ended, the information contained in columns 13 to 16 were to be given in columns 11 and 12 during the visits when such periods have come to an end.

3.26. The history of all past terminations relating to every 'ever-married' woman in block 6 was recorded in block 9 together with certain demographic characteristics of the



woman and her husband (living or dead). This block was filled up only in the last visit to the household since it was considered that in making such enquiries about the past histories of the woman a more intimate acquaintance with the household was desirable to enlist the full co-operation of the household. The information on the age of the mother at successive terminations and the result of each termination were entered in this block. In the event of the last termination of any woman taking place within one year prior to the visit, the month of birth and the result of the termination were to be recorded in columns 32 and 33 respectively. This was necessary because for such children who have not completed one year of age the period of exposure had to be separately estimated. Also in respect of such a termination, information on ante-natal care was to be collected and entered in column 34. A two-digit composite code, the left-hand digit indicating the type of attendance and the right-hand digit indicating the number of such attendances was to be used.

## CHAPTER 4

### INFANT MORTALITY

4.1. One of the vital rates with a very wide range of applicability in the field of public health is the infant mortality rate. This is expressed as the number of deaths under one year per thousand live births. Apart from its practical utility in maternal and child health studies, its value as a general index of the health of a population group is in no way inferior to such mortality rates as standardized or life table death rates. Its high sensitivity to the general living conditions of the population to which it relates makes it an immensely valuable measure for comparison of health conditions of different population groups.

4.2. For the purpose of this survey, a household which has been selected as the ultimate sample unit is defined as a group of persons living together for a period of one month previous to the date of visit. Obviously, this is too short a period for studies of vital events like deaths, etc. This definition necessarily excludes from its purview the consideration of vital events relating to persons who were members of the household earlier but not at the time of survey. But the scope of the definition of the household had to be restrained in the above manner to obviate certain practical complications that might arise due to the mobility of the members of households. For example, it is not unlikely that with the death of the principal earner of a household his dependents are eventually absorbed as members of other households leading to a dissolution of the original household in which the death occurred. Such circumstances will naturally lead to gross under-estimation of deaths since the sample frame excludes households in which such events occurred. In this survey, however, a continuous record of births, deaths and illnesses occurring to the members of the selected households during a span of three months could be maintained since the survey was conducted by four visits at specified intervals. Even a period of three months, it is to be admitted, falls short of the requirements for obtaining sufficiently reliable estimates of the age-specific mortality rates which enter into the computation of standardized or life-table death rates.

4.3. Since married women belonging to the selected households formed a representative sample of all living married women in the population, the infant mortality rates relating to the live births which occurred to them during the past one or even two years could be



considered as adequately depicting the infant mortality conditions prevailing during the period. This is true because all children born during the period except those born to women dying during the period will be represented in the sample whether or not such infants could be considered as members of the household according to the definition. The fraction of women dying during the period of one or even two years being negligible, the infant mortality rate calculated on the basis of the surviving women is not likely to be significantly different from the one calculated by inclusion of the dying women.

4.4. The above considerations go to show that the infant mortality rate, apart from its usefulness as a general mortality index, has a definite advantage over other general mortality rates with regard to the statistical validity when the study is based on sample survey data. For this reason, the infant mortality rate has been utilized in this study as a general mortality index for the comparison of different social and economic classes. All live births occurring to each married woman of the household, the order of such births together with information on survival at the end of one year after birth can be obtained from block 9 of the schedule. For the computation of infant mortality rates of different population groups, a short reference period of two or even five years would have been desirable because such a period has two advantages, namely, (i) infant mortality rate based on recent live births would be more appropriate to depict the recent health conditions of the population groups and (ii) the infant mortality rate would be statistically more valid as it would be relatively free from recall lapse. Due to the limitations in the present data, the choice of such a short reference period becomes almost impracticable and a much wider basis to include all live births that occurred upto the date of survey to each ever married woman was resorted to. The validity of the comparison on the basis of infant mortality rates thus obtained may be disputed as it was pointed out in the National Sample Survey Report No. 7 that the infant mortality rates relating to the pre-1930 marriage cohorts were inordinately low compared to the official rates for corresponding periods. As the NSS data were collected in the year 1952, it might be expected that about 40 per cent of the pre-1930 marriage cohorts might have died earlier to the date of survey with the result that this group got automatically excluded from the analysis. Since the cohorts thus excluded were likely to belong to the high mortality group, the infant mortality rate estimated from the surviving group might have been probably biased towards a lower value. Under the circumstances, it is difficult to attribute the entire difference between the NSS rate and the official rate to recall lapse.

4.5. In the following analysis, therefore, we have, in the first instance, limited our study exclusively to the group of ever-married women with at least five terminations. The live-births occurring to such women were arranged by parity and the infant mortality rate for each successive parity was estimated. The rates for the rural and urban samples are shown in Table 4.1.

4.6. The results given in the above table clearly show, as should be expected, that the infant mortality rates relating to the initial parities, particularly to the first two, are substantially higher than those recorded for subsequent parities. If recall lapse did really effect the infant mortality rates in the distant past, then the estimated rates for the initial two parities could not have exceeded those for subsequent parities both among the rural and urban populations to the observed extent. Moreover, since the infant mortality rates for the first five parities are based on births relating to the same group of women, and if



it can be assumed that the average time interval between the first and fifth parities is about 15 years, there seems to be not much ground to suspect a substantial reduction in the infant

TABLE 4.1. MORTALITY RATES FOR INFANTS BORN TO WOMEN HAVING 5 OR MORE TERMINATIONS

order of birth	rural	urban
(1)	(2)	(3)
1. 1st	244.00 (500) <sup>1</sup>	205.13 (156)
2. 2nd	228.00 (500)	202.53 (158)
3. 3rd	198.02 (505)	121.02 (157)
4. 4th	157.37 (502)	132.08 (159)
5. 5th	134.92 (504)	87.50 (160)
6. 6th	118.13 (364)	123.81 (105)
7. 7th	101.21 (247)	51.28 (78)
8. 8th & above	96.67 (300)	148.15 (108)
9. all orders	169.49 (3422)	139.69 (1081)

<sup>1</sup> Figures in parentheses refer to the numbers of live-births on which the infant mortality rates are based.

mortality rate due to recall lapse. In parities higher than the fifth, there would be a successive reduction from the initial set of women as some of them with fewer terminations are likely to drop out and as such their rates are not strictly comparable with the rates for earlier parities.

4.7. As the parity advances, the estimated infant mortality rates also correspond more and more closely to recent events and if recall lapse were significant there should have been a tendency for the estimates to rise with advancing parities. The estimates shown in Table 4.1, however, do not give any evidence of a rise.

4.8. If, on the other hand, a large fraction of the class of women included in the foregoing analysis were too advanced in age (say, 60 years or above), then even in the later parities, one should expect a number of births that took place so long ago as to be affected by recall lapse. It may, therefore, be argued that such comparisons as made above might hardly detect the existence of recall lapse unless the rates for advanced parities really relate to very recent events. As such, the births occurring to ever-married women of age 43 years and over were classified into two chronological groups, births occurring within



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

and earlier to the period of 15 years preceding the date of survey, and the infant mortality rates for these two groups were mutually compared. The results are given in Table 4.2.

TABLE 4.2. MORTALITY RATES FOR INFANTS BORN TO EVER  
MARRIED WOMEN AGED 43 YEARS OR OVER  
ACCORDING TO CHRONOLOGICAL GROUPS

period when births occurred	infant mortality rate	
	rural	urban
(1)	(2)	(3)
within 15 years preceding the date of survey	69.26 (231) <sup>1</sup>	119.05 (84)
15 or more years earlier to survey	160.28 (1984)	137.17 (678)

<sup>1</sup> Figures in parentheses refer to the numbers of live births on which the infant mortality rates are based.

4.9. Since the earlier parities are associated with higher infant mortality rates than the latter ones, the chronological comparison should have been attempted at corresponding parities. As this was impracticable with the data in hand, all parities were mixed together into one lot within each chronological group. As the recent chronological group is likely to be more heavily loaded with later parities one should naturally expect a lower infant mortality among them. Among rural births this is clearly indicated by the rates entered in Table 4.2. If recall lapse was operative, such striking difference in the infant mortality rates would not have been observed. As most of the women included in the urban sample were of advanced age, hardly 84 births occurred to them during the last 15 years and the remaining 678 births were classified in the older chronological group which naturally included a substantial number of later parities as well. The contrast between the infant mortality rates for the two chronological groups is, therefore, relatively less marked than that observed for the rural samples. In any case, it appears from the above analysis that the effect of recall lapse is not statistically very significant.

4.10. In this study, the estimates of infant mortality rates have been based on all live-births that occurred to ever-married women of the respective population groups upto the date of survey. The infant mortality rates have been estimated for the rural and urban sectors separately and the results are shown for each of the two sub-samples in Table 01.1 of Appendix 1.

4.11. In what follows, an attempt is made to study the association of infant mortality rate with such factors as nutritional level of the household, housing condition, educational and occupational status of fathers.

4.12. *Nutrition.* Nutrition being a vital factor of health, a separate block (block 4) was devoted in the schedule for entering the various items regarding the dietary composition of the households. The scope of the survey did not, however, permit a detailed study of the various items comprising the diet and assess the nutritional level of the household based on its departure from an ideal or balanced diet. It is well-known that the majority of the Indian households can hardly avail even the basic energy-giving food



articles like cereals, etc., for satisfying their hunger. This being the case, differentiation of diet can be effected even on the basis of the quantity of cereals consumed. If appropriate adult equivalents for various age and occupational groups were available, one could have classified the households by varying degrees of nutritional level on the basis of the quantity of cereals consumed. In this study, however, it has been assumed that a household which avails even a small quantity of milk, meat, fish, fruits, etc., for whatever it is worth, must be doing so only after satisfying its basic needs in respect of cereals. On this assumption, diets which are almost completely devoid of milk, fish, meat etc., have been placed in the category 'low level of nutrition' and the remaining diets in the category 'high level of nutrition'. The infant mortality rates observed in the above two dietary classes are shown separately for rural and urban sectors in Table 4.3.

TABLE 4.3. INFANT MORTALITY RATE ACCORDING TO LEVEL OF NUTRITION IN WEST BENGAL

nutritional level	rural		urban	
	no. of live births	infant mortality rate	no. of live births	infant mortality rate
(1)	(2)	(3)	(4)	(5)
1. high	155	116.13	659	110.77
2. low	5384	170.69	1186	150.08
3. total	5539	169.16	1845	136.04

4.13. In both the rural and urban sectors, the differences in the recorded infant mortality rates are significant, the difference being more pronounced in the case of the rural group. Inasmuch as the rate corresponding to the nutritional group classed as 'high' in the rural sector is based on an inadequate number of live births, the observed difference is to be accepted with a little caution.

4.14. *Housing.* Housing is an important factor affecting the health of a population. One of the objectives of this survey was to study the association between the sanitary assessment of the household and certain important health indices like infant mortality rate and thereby to evolve a suitable methodology for collecting information on housing and environmental conditions. Hence, a few questions relating to the housing and sanitary conditions of the household and its environment were included in block 5 of the schedule. Only such aspects which were more easily definable and less subjective in nature have been introduced in this block and as such it had to be necessarily short. Nevertheless, the survey revealed some shortcomings in the method of approach adopted, particularly, with reference to the rural sample.

4.15. In the rural area, most of the households were without latrines and as such no variation in the type of latrine used was ascertainable from the returns. The information obtained on ventilation of households though somewhat subjective in nature, proved to be defective for the rural and urban areas on account of the vagaries of the investigators. As regards the general sanitation the investigators relied more on the relative differences among the households allotted rather than on the objective classifications specified for the investigational procedure. In the rural sector particularly, where the village as a whole was



allocated to each investigator, the entries relating to this aspect for all the households were almost identical, the nature of such entries being determined largely by his personal impressions. In towns and cities, however, the households allotted to each investigator being situated over an area within which conspicuous differences in sanitary conditions existed, there was a greater degree of variation in the general sanitation codes entered by him. With regard to source of drinking water, taps, tubewells and ordinary wells were almost universal in the urban areas, whereas in the rural areas the majority of the people depended on wells or ponds. Hence, with the type of information available, it was thought that no useful classification of rural households could be possible on the basis of such characteristics as sanitary conditions of the dwelling place. For the classification of the urban households, information on latrine and general sanitation of surroundings alone could provide a valid basis. The households with code 1 for both 'latrine' and 'general sanitation' were classified as 'housing good' and the rest as 'housing bad'. The infant mortality rates in these two classes of the urban population are shown in Table 4.4 which clearly indicates that the population which avails better sanitary amenities in and around its dwelling place is associated with a lower infant mortality rate.

TABLE 4.4. INFANT MORTALITY RATE ACCORDING TO HOUSING CONDITIONS IN WEST BENGAL (URBAN)

housing condition	live births	infant mortality rate
(1)	(2)	(3)
1. good	366	84.70
2. bad	1479	148.75
3. total	1845	136.04

4.16. *Education.* No doubt, the two criteria considered above, namely, nutrition and sanitary condition are useful indices to study the state of health of a community. Data have also been collected in this survey on the literacy and occupational status of the population. Though these may not directly influence the health of the people, in earlier studies it has been noticed that such socio-economic factors as education and occupation bear a strong association with the infant mortality rate. Hence, these socio-economic factors can be used with great advantage in health studies for more effective stratification of the households for sample selection.

4.17. In so far as education improves the quality of maternal care and personal hygiene by raising the level of health consciousness of the community its association with infant mortality rate may be regarded as a direct one. Besides the above, its indirect relationship through such health factors as nutrition, sanitary condition etc., enhances the degree of such association. For a critical evaluation of the importance of education it would have been desirable to consider the education of mothers but under the prevailing circumstances this is not feasible as the large majority of women especially in the rural sector are illiterate. Hence, it was decided to take into account the education of fathers for the above analysis. Here again, due to the limited scope of the data only two broad categories were considered, namely (i) births relating to couples in which the male partners had matric or



higher qualifications and (ii) the remaining births relating to couples in which the male partners were either illiterate or under-matrices. The proportion of births belonging to the higher educational group (matric or above) to total births were roughly 22% in the urban and 2% in the rural sector. The estimated infant mortality rates for the two literacy classes are shown separately for rural and urban samples in Table 4.5.

TABLE 4.5. INFANT MORTALITY RATES ACCORDING TO LITERACY STATUS OF FATHERS IN WEST BENGAL

literacy status	rural		urban	
	live births	infant mortality rate	live births	infant mortality rate
(1)	(2)	(3)	(4)	(5)
1. matric and above	108	55.56	404	76.73
2. below matric including illiterates	5431	171.42	1441	152.67
3. total	5539	169.16	1845	136.04

4.18. The results are indeed very striking, the higher literacy group showing an infant mortality rate which is nearly one-third and one-half of the lower literacy group in the rural and urban areas respectively. This is a clear indication that literacy status is an excellent criterion for stratification of urban households in sample surveys of this nature. In view of the fact that only about 2% of the births in the rural areas correspond to the higher literacy group, a similar stratification is of doubtful utility in studies of this kind for rural populations.

4.19. *Occupation.* In order to study the behaviour of infant mortality rates in different occupational groups a similar analysis as above was carried out. Here again, the occupational classification had to be confined to those of the male partners only. Due to paucity of data the analysis had to be limited to four broad categories of occupational status for the urban and rural sectors as indicated below.

- 4.20. *Urban :*
1. Manual labour (mostly unskilled industrial labour, domestic servant, porter, hawker, rickshaw puller, artisan, etc.)
  2. Lower professions and inferior business (clerk, school teacher, retail trader, shop assistant, skilled industrial labour, etc.)
  3. Higher professions and superior business (doctor, professor, engineer, lawyer, wholesale trader, etc.)
  4. Non-gainful occupations (rent receiver, remittance receiver, beggar, etc.)

- 4.21. *Rural :*
1. Agricultural and other rural labour (landless agricultural labour, artisan, fisherman, cooly, etc.)
  2. Agricultural operations (cultivator owning land, share-cropper, etc.)
  3. Professions and trade (teacher, doctor, priest, retail trader, etc.)
  4. Non-gainful occupations (rent receiver, remittance receiver, etc.)



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

4.22. In both the urban and rural populations, the class 'non-gainful occupations' contains a highly heterogeneous social group as it includes all persons returned as 'not in the labour force', irrespective of their living standards.

4.23. The estimated infant mortality rates for the different occupational classes are shown in Tables 4.6 and 4.7 for the urban and rural populations respectively.

TABLE 4.6. INFANT MORTALITY RATE ACCORDING TO OCCUPATION OF FATHERS  
IN WEST BENGAL (URBAN)

occupation class	no. of live births	infant mortality rate
(1)	(2)	(3)
1. manual labour	492	170.73
2. lower professions and inferior business	1061	126.30
3. higher professions and superior business	155	45.16
4. non-gainful occupations	137	189.78
5. total	1845	136.04

TABLE 4.7. INFANT MORTALITY RATE ACCORDING TO OCCUPATION OF FATHERS  
IN WEST BENGAL (RURAL)

occupation class	no. of live births	infant mortality rate
(1)	(2)	(3)
1. agriculture and other rural labour	1486	154.10
2. agricultural operations	3120	183.33
3. professions and trade	429	158.51
4. non-gainful occupations	504	134.92
5. total	5539	169.16

4.24. In the lowest occupation class (manual labour) of the urban population the recorded infant mortality rate is as high as 170.73 per 1000 live-births, whereas in the highest social class (higher professions and superior business) it is as low as 45.16 per 1000 live-births and that for the intermediate class (lower professions and inferior business) it is 126.30 per 1000 live-births. From these results, it is quite apparent that the infant mortality rates tend to decrease as one goes up the social ladder. In rural areas, however, due to the inadequate number of sample households belonging to the higher professional group, they were merged with the lower professions to form class 3 (professions and trade). Due to the preponderance of the households belonging to the lower professions, infant mortality rate observed for this combined group was considerably enhanced. As regards social class 2 (agricultural operations), which includes every cultivator owning land, however small his holdings may be, and every share-cropper, howsoever small the area operated by him may be, is certainly a heterogeneous group. For these reasons, the results



in Table 4.7 do not suggest any occupational differentials in the infant mortality rate. This, perhaps, indicates that for stratifying the rural population into social classes, it may be desirable to take cognizance of certain other relevant factors. May be, social differentiations based on either income or land operated or owned may be more appropriate for health studies as this will more closely correspond to the actual living standards of the households.

4.25. In this context, it may be appropriate to consider the Registrar General's figures of infant mortality rate by social class of father in England and Wales in 1939 (Table 4.8) and see how the social position of the community affects the infant mortality rate.

TABLE 4.8. INFANT MORTALITY RATE ACCORDING TO SOCIAL CLASS OF FATHER  
IN ENGLAND AND WALES IN 1939

social class		infant mortality rate
(1)		(2)
1. class I—the professions, commissioned officers and well-to-do people concerned with finance, shipping etc.		26.8
2. class II—intermediate between class I and skilled workers		34.4
3. class III—skilled workers		44.4
4. class IV—intermediate between skilled and unskilled workers		51.4
5. class V—unskilled workers		60.1

4.26. The above figures exhibit a high degree of consistency and regularity in the changing pattern of infant mortality rate with changing social class. A similar feature is observed in the occupational classes in urban West Bengal. This clearly suggests that occupational or social status offers an excellent criterion for a more effective stratification of the urban population for sample selection in studies of this kind.

4.27. In conclusion, it may be stated that the higher level of nutrition and the higher level of literacy and occupational status are associated with lower infant mortality rates. It is also interesting to note from the results given above that when comparable groups are matched against each other the rural groups generally show higher infant mortality rates than their urban counterparts. However, the infant mortality differentials estimated from registration data indicate an entirely opposite picture. Even in respect of the overall estimate of infant mortality rate the value observed in this study for the rural sector appears to be higher than that observed for the urban sector. If this is true, it is obviously at variance with the accepted notion on urban-rural differentials based on registration data. Since official figures do not relate to allocated rates, further examination of the registration data is necessary before arriving at any definite conclusions.

4.28. It is a known fact that infants die at a faster rate during the earlier periods of their life. It can be reasonably assumed that in rural West Bengal about 40 per cent of the infants die before they complete one week. Possibly, the infant mortality rate observed in case of rural births may be largely attributable to deaths occurring during this stage of life due to inadequate and unsatisfactory maternity aid available to rural mothers.



## CHAPTER 5

## MORBIDITY

5.1. At present the statistical data collected on health aspects are so meagre and unreliable in our country that they can utmost provide a very crude and hazy outline of the health conditions of the nation. But these are not enough for sound public health administration which obviously has to be based on reliable and adequate factual data. In the past, a few special surveys relating to malaria, tuberculosis, leprosy, etc., have been attempted in selected areas only. Such surveys are useful, no doubt, in shaping the public health policy to some extent in these areas, but in order to plan for the improvement of health standards of the community as a whole, an appraisal of the disease and medical care pattern is essential and this can be done only by a general health survey.

5.2. The first general health survey to be conducted in this country was the Singur Health Survey (Lal and Seal, 1949). But this study, though comprehensive in nature, had to be necessarily confined to a small rural area in West Bengal comprising of the 4 union boards falling within the sphere of operation of the Singur Health Centre. It was expected that when the report of the above study was published, similar enquiries would be made in other parts of India to obtain a general picture of the morbidity and medical care pattern. But till the year 1955 no attempts were known to have been made in this direction.

5.3. The lack of initiative in sponsoring surveys of this kind certainly reflects the peculiar difficulties inherent in such surveys, especially in countries like India where more experiments in survey technique and procedure remain to be done before pushing through full-fledged health surveys on a vast scale. The West Bengal Health Survey, as has already been pointed out, is only such an experiment which was undertaken mainly for the purpose of developing a methodology for the collection of health and medical care statistics.

5.4. Certain experienced public health workers with whom the scheme of this survey was discussed, were rather critical of the approach that had been adopted in this survey. They raised pertinent questions regarding the validity of the morbidity statistics collected by such surveys relying mainly on non-medical investigators. The most serious objection centred round the correctness of the classification of diseases by causes. It may be mentioned that even in countries where medical care has become almost universal, the competency of the informants, generally the heads of households, to report the true cause of the disease which has become a thing of the past, is somewhat questionable. More reliance is, therefore, placed on prevalence rates for certain chronic diseases obtained by complete physical examination of the selected individuals including laboratory tests by trained medical personnel. Even if such a scheme were possible on a nation-wide scale for a country like India having only meagre resources, it could cover only some of the important chronic diseases and the question regarding the incidence of acute diseases which form a substantial bulk of the total morbidity of the country will still remain unsolved.

5.5. One of the chief objectives of this study, therefore, was to assess the extent of agreement between the reported causes of diseases and the diagnostic reports at the time of treatment wherever such reports were accessible to the investigators. Though the investigators were briefed to avail of the medical diagnostic reports wherever they were available, it was found that only 4 cases of illness were accompanied by such reports in spite



of the fact that about two-third of the cases were medically treated. Presumably, the odds against collecting such information were so great that the investigators could not possibly succeed in carrying out the instructions. Of the 4 cases for which medical confirmation was available, only 2 cases—one of 'appendicitis' and the other of 'pneumonia'—were found to be in complete agreement with the investigators' returns. For the remaining two cases which were medically declared as tuberculosis (pulmonary), the informants returned them as merely 'cough and fever'.

5.6. Though in a substantial number of cases, certain remarks regarding the nature of the diseases made by the informants of their own accord and entered in the 'remarks' column of block 7, were very helpful to the medical experts in arriving at a proper classification of the diseases, it has to be admitted that due to non-availability of confirmatory evidence from the attending physicians, no check on the validity of the returns could be made.

5.7. As the question of validity is an important one on which depends to a considerable degree the success of a morbidity study, a 'Validity Survey' was initiated in 1956. For the purpose of this survey, two teams of investigators, one medical and the other non-medical, were employed. The medical investigators were medical graduates with considerable professional experience. The non-medical investigators, on the other hand, did not have any particular training or knowledge in public health. Names of patients with addresses were collected from the medical out-patient department of the R.G. Kar Medical College Hospitals, which is one of the leading hospitals in Calcutta. These names were supplied to a set of pilot investigators for verifying the addresses as well as to note down the names of all members of the households. After this preliminary listing was done, the medical and non-medical investigators were given the names and addresses of the heads of the households. They were also given the names of all the members of the households to ensure that the particular person who had been to the hospital and about whose illness information was available, was not omitted from investigation. They were required to collect details about all illnesses occurring to the members of the households within a reference period of one month. The non-medical investigators were instructed to put down the cause of the disease as stated by the informants and supplement them with details of signs, symptoms etc., of the disease, if such information was forthcoming from them on their own initiative. The medical investigators, on the other hand, had a greater degree of freedom in that they could interrogate the heads of the households or the patients themselves for their views on the disease. This freedom was not allowed to the non-medical investigators because it was thought that the non-medical investigator by virtue of his not having any medical knowledge or training was not competent to suggest leading questions to arrive at a proper conclusion as regards the exact nature of the disease. No indication whatsoever was given to the investigators as to the source of these addresses. When these households have been contacted and necessary information gathered in the schedules specially designed for the purpose, the returns were compared with the hospital diagnoses. It would seem that the best way of doing this would have been to send both the types of investigators to the same households and compare their results. But this procedure did not appeal to us as in such a short time which was generally less than a month, it was not desirable to subject a household to a series of questions by different investigators, especially when a disease was prevailing in that household. Moreover, there might be a tendency for the first investigation to influence the result of the second.



# A PILOT HEALTH SURVEY IN WEST BENGAL—1955

5.8. In order to make an overall assessment of the relative merits of the two types of investigating teams, the cases diagnosed in each class of diseases were equally apportioned between the two teams. It could be seen that by the above arrangement if the extent of misclassification varied with the nature of disease, the odds were equally balanced between the two teams. A total of 396 cases could be contacted in their households and of these 198 were investigated by the medical investigators and the remaining 198 by the non-medical investigators. The investigator's returns were then compared with the corresponding reports obtained from the hospital register and the results of this comparison are shown in Tables 5.1 and 5.2 for medical and non-medical investigators. It may be mentioned here that the non-medical investigators' reporting of diseases being sometimes vague, they were allocated to proper disease groups by a panel of medical experts on the basis of signs, symptoms and other available particulars of such diseases.

TABLE 5.1. THE EXTENT OF AGREEMENT BETWEEN THE HOSPITAL DIAGNOSES AND RETURNS OF THE MEDICAL INVESTIGATORS

disease	complete agreement	no agreement	doubtful	not recorded	total
(1)	(2)	(3)	(4)	(5)	(6)
1. group I—tuberculosis (pulmonary)	—	6	—	—	6
2. group II—malaria	3	10	5	2	20
3. group III—dysentery	4	9	2	—	15
4. group IV—other infectious and parasitic diseases	6	5	—	—	11
5. group V—allergic, endocrine system, metabolic and nutritional diseases	7	2	3	—	12
6. group VI—diseases of blood and blood-forming organs	1	7	2	—	10
7. group VII—mental, psychoneurotic and personality disorders and diseases of the nervous system and sense organs	2	6	3	—	11
8. group VIII—diseases of the circulatory system	2	2	1	1	6
9. group IX—influenza	7	5	3	2	17
10. group X—bronchitis	11	10	6	1	28
11. group XI—other respiratory diseases	1	8	3	—	12
12. group XII—diseases of digestive system	18	15	2	4	39
13. group XIII—diseases of digestive system	—	2	—	—	2
14. group XIV—diseases of genito-urinary system	—	—	—	—	—
15. group XV—diseases of bones and organs of movement	4	3	—	1	8
16. total	—	—	—	—	—
15. group XV—other diseases	67	90	30	11	198



TABLE 5.2. THE EXTENT OF AGREEMENT BETWEEN THE HOSPITAL DIAGNOSES AND RETURNS OF THE *NON-MEDICAL* INVESTIGATORS

disease	complete agreement	no agreement	doubtful	not recorded	total
(1)	(2)	(3)	(4)	(5)	(6)
1. group I—tuberculosis (pulmonary)	3	3	—	—	6
2. group II—malaria	6	10	3	2	21
3. group III—dysentery	1	12	—	2	15
4. group IV—other infectious and parasitic diseases	—	10	1	—	11
5. group V—allergic, endocrine system, metabolic and nutritonal diseases	2	9	—	1	12
6. group VI—diseases of blood and blood-forming organs	1	7	—	—	8
7. group VII—mental, psychoneurotic and personality disorders and diseases of nervous system and sense organs	2	9	—	—	11
8. group VIII—diseases of the circulatory system	2	2	—	1	5
9. group IX—influenza	5	6	1	5	17
10. group X—bronchitis	6	17	3	4	30
11. group XI—other respiratory diseases	3	11	—	1	15
12. group XII—diseases of digestive system	26	8	—	5	39
13. group XIII—diseases of genito-urinary system	—	—	—	—	—
14. group XIV—diseases of bones and organs of movement	5	2	—	1	8
15. group XV—other diseases	—	—	—	—	—
16. total	62	106	8	22	198

5.9. Two types of discrepancies are possible in this situation. The first is that the diagnosis entered in the hospital register does not tally with those obtained from the investigators' reports and secondly, certain individuals who were known to have attended the hospital in connection with certain definite illness were not reported by the investigators as having suffered from any illness during the reference period. The first type of discrepancy, therefore, is ascribable to misclassification of either the hospital or the investigator or by both. The second type of discrepancy is certainly an error ascribable to the method of investigation probably due to recall lapse. Occasionally, a person who attended the hospital in connection with a certain illness might have been ill due to another illness also during the reference period. In such a situation if the investigators' reports did not tally with the hospital reports, it would not be possible to state clearly whether it was a misclassification or an omission of the disease for which the enquiry was made. If, however, the date of onset of any disease reported by the investigators preceded the date of hospital attendance it was assumed that the report related to the disease for which hospital aid was sought. On the other hand, if the date of onset reported was later than the date of hospital attendance, it was difficult to decide whether the investigators' report related to the same disease as was treated in the hospital or to a different one. In the latter situation the discrepancies were



# A PILOT HEALTH SURVEY IN WEST BENGAL—1955

assigned in a separate column 'doubtful' in Tables 5.1 and 5.2. Further, there were certain cases where there was complete agreement between the investigators' reports and the corresponding entries in the hospital register, but the date of onset reported by the investigators was subsequent to the date of hospital attendance. But as these cases were either of a chronic or intermittent nature, it could be reasonably assumed that the investigators' reports related to the same diseases as were treated in the hospital. It may be seen from Tables 5.1 and 5.2 that out of 198 cases of illness 30 and 8 cases allotted to the medical and non-medical investigators respectively did not tally with the hospital entries for lack of knowledge, whether the investigators' reports related to the corresponding diseases for which the enquiry was made. Out of 168 cases for which medical investigators' reports could be tallied with the corresponding hospital entries, 11 were missed and 90 were misclassified, whereas for the non-medical investigators, out of 190 cases for which the reports could be tallied with the hospital entries 22 were missed and 106 were misclassified. The percentage of cases missed by the medical investigators is only 6.5 per cent as compared with 11.6 per cent missed by the non-medical investigators. This indicates that there will be more response obtained from the informants if medical investigators are employed. Among 157 cases reported by the medical investigators which could be tallied with hospital entries, 90 cases or about 57 per cent were misclassified, whereas among 168 similar cases reported by the non-medical investigators, 106 cases or about 63 per cent were misclassified. The above results suggest that both in respect of extent of response from the informants as well as in the extent of correct classification, the performance of the medical investigators seems to be slightly more satisfactory than that of the non-medical team. The extent of misclassification in the 15 groups of diseases individually are shown in table below.

TABLE 5.3. PERCENTAGE MISCLASSIFICATION AMONG THE MEDICAL AND NON-MEDICAL INVESTIGATORS IN DIFFERENT DISEASE CATEGORIES

disease	medical investigator	non-medical investigator
(1)	(2)	(3)
1. group I—tuberculosis (pulmonary)	100.0	50.0
2. group II—malaria	76.9	62.5
3. group III—dysentery	69.2	92.3
4. group IV—other infective and parasitic diseases	45.5	100.0
5. group V—allergic, endocrine system, metabolic and nutritional diseases	22.2	81.8
6. group VI—diseases of blood and blood-forming organs	87.5	87.5
7. group VII—mental, psychoneurotic and personality disorders and diseases of nervous system and sense organs	75.0	81.8
8. group VIII—diseases of circulatory system	50.0	50.0
9. group IX—influenza	41.7	54.5
10. group X—bronchitis	47.6	73.9
11. group XI—other respiratory diseases	88.9	78.6
12. group XII—diseases of digestive system	45.5	23.5
13. group XIII—diseases of genito-urinary system	100.0	—
14. group XIV—diseases of bones and organs of movement	42.9	28.6
15. group XV—other diseases	0.0	—
16. total	57.3	63.1



5.10. That the extent of disagreement in the reporting of the diseases varies with the type of diseases investigated is quite evident from Table 5.3. Moreover, the above table singles out such types of diseases which are likely to be more often misreported by the medical and non-medical investigating teams as well as such diseases for which the extent of agreement with the corresponding hospital diagnosis does not show conspicuous difference between the two investigating teams. For instance, diseases belonging to groups 6 (diseases of blood and blood-forming organs), 7 (mental, psychoneurotic and personality disorders and diseases of the nervous system and sense organs) and 8 (diseases of the circulatory system) show almost equal tendency to be misclassified by the medical and the non-medical investigators. On the other hand, diseases belonging to groups 1 (pulmonary tuberculosis), 2 (malaria), 11 (other respiratory diseases), 12 (diseases of the digestive system) and 14 (diseases of bones and organs of movement) are generally misreported to a greater extent by the medical investigators and diseases belonging to groups 3 (dysentery), 4 (other infective and parasitic diseases), 5 (allergic, endocrine system, metabolic and nutritional diseases), 9 (influenza) and 10 (bronchitis) are similarly misclassified by the non-medical investigators. These results are important in as much as they are suggestive of the quality of reporting by the medical and non-medical investigating teams with respect to different disease groups. In order to properly assess the validity of the rates obtained by the two types of investigators and to investigate the nature of misclassification, further examination of the data is essential. Tables 5.4 and 5.5 give the two-way comparison of the reports of the investigators with the corresponding entries in the hospital register.

5.11. In the above tables the diagonal entries represent those cases where there is complete agreement between investigators' returns and the hospital diagnoses. The figures in rows indicate the extent of misclassification occurring for each type of disease taken from the hospital register. If we assume for the purpose of argument that the reports obtained from the hospital register are correct as they were made during the time of treatment when the disease prevailed, the entries in each row divergent from the diagonal cell will show the extent of under-estimation of the morbidity rate for this particular group of diseases due to misclassification. On the other hand, the entries in any column diverging from the diagonal cell will indicate the extent of over-estimation of the morbidity rate due to inclusion of diseases of other categories in this group by the investigators. Of course, this assumption regarding the hospital diagnoses need not be true and, therefore, this study of the divergence between the two types of classification can be interpreted only as a lack of agreement and no further.

5.12. The direction in which the misreporting takes place and the ultimate effect of such a misclassification on the morbidity rates from different diseases are points which deserve consideration. In the following paragraphs an attempt is made to discuss briefly the net results arising out of misreporting of diseases.

5.13. There is a general tendency noticeable for pulmonary tuberculosis to be invariably misclassified as either asthma or bronchitis. Other forms of tuberculosis are also usually misreported. It is also found that the non-medical investigators have marked tendency to report non-tuberculosis cases as pulmonary tuberculosis cases resulting in an exaggeration of the pulmonary tuberculosis rate estimated from their returns.

5.14. Malaria is sometimes reported as influenza or other respiratory diseases or diseases of the digestive system and non-malaria cases are not generally returned as malaria



# A PILOT HEALTH SURVEY IN WEST BENGAL—1955

TABLE 5.4. COMPARISON OF THE MEDICAL INVESTIGATORS' RETURNS WITH THE CORRESPONDING HOSPITAL ENTRIES  
(investigator)

disease	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
1. group 1—tuberculosis (pulm.)	.	.	.	.	.	1	.	.	.	.	2	1	.	.	1	1	.	6	.	.	6
2. group 2—malaria	.	3	.	1	.	.	1	1	1	2	1	2	1	.	.	2	.	13	5	2	20
3. group 3—dysentery	.	.	.	4	6	.	1	.	1	1	1	1	4	1	1	.	.	13	2	.	15
4. group 4—other infective and parasitic diseases	.	.	.	.	.	7	1	.	.	.	1	.	.	.	2	1	.	11	1	.	11
5. group 5—allergic, endo, meta, nut. disorders	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	9	3	.	12
6. group 6—diseases of blood and blood-forming org.	.	.	.	1	.	.	1	1	1	.	.	1	3	3	.	.	.	8	2	.	10
7. group 7—mental, psycho-neurotic and nervous system	.	.	.	.	.	1	1	2	.	.	.	.	.	2	2	.	.	8	3	.	11
8. group 8—diseases of circulatory system	.	.	.	.	.	.	.	.	2	.	1	.	.	.	1	.	.	4	1	1	6
9. group 9—influenza	.	.	1	.	.	.	.	.	.	7	1	2	.	.	1	.	.	12	3	2	17
10. group 10—bronchitis	.	.	.	.	.	1	.	1	1	4	11	1	3	1	.	.	.	21	6	1	28
11. group 11—other respiratory diseases	.	.	.	.	.	1	.	1	.	1	3	1	.	1	.	1	.	9	3	.	12
12. group 12—diseases of digestive system	1	.	.	4	.	.	.	.	.	2	.	.	18	5	1	2	.	33	2	4	39
13. group 13—diseases of genito urinary sys.	.	.	.	1	.	.	.	.	.	.	.	.	1	.	.	.	.	2	.	.	2
14. group 14—diseases of bones and organs of movement	.	.	.	.	1	.	.	1	.	1	.	.	.	.	4	.	.	7	.	1	8
15. group 15—other diseases	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
16. total	1	4	11	7	11	5	5	4	18	20	8	30	12	13	8	1	157	30	11	198	



TABLE 5.5. COMPARISON OF THE NON-MEDICAL INVESTIGATORS' RETURNS WITH THE CORRESPONDING HOSPITAL ENTRIES

diseases	(investigator)																				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
1. group 1—tuberculosis (pulm)	3					1							2					6			6
2. group 2—malaria	1	6		1	1		1	3					2	1				16	3	2	21
3. group 3—dysentery	1	1	1	1	1				1				7		1			13		2	15
4. group 4—other infective and parasitic diseases						1		2		1	1	1	2		2			10	1		11
5. group 5—allergic, endocrine, met. and nut. diseases					1	2		1		1	2	1	1			1	1	11		1	12
6. group 6—diseases of blood and blood-forming organs						1	1			1			2	1		1	1	8			8
7. group 7—mental, psychoneurotic and nervous systems	1							2					3	2	2		1	11			11
8. group 8—diseases of circulatory system									2				1		1			4		1	5
9. group 9—influenza	1			1						5		1		1				11	1	5	17
10. group 10—bronchitis	2				2	2				4	6	1	3		2	1		23	3	4	30
11. group 11—other respiratory diseases	1						1			2	3	3	3	1				14		1	15
12. group 12—diseases of the digestive system					2	2		1					26	3				34		5	39
13. group 13—diseases of genito-urinary system																					
14. group 14—diseases of bones and organs of movement													1		5		1	7		1	8
15. group 15—other diseases																					
16. total	10	7	3	3	7	9	3	9	3	14	12	7	53	9	13	5	4	168	8	22	198



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

cases by both the medical and non-medical investigators. Hence it seems that the malaria rate as obtained by a household canvass is likely to be an under-estimate.

5.15. Dysentery cases are usually misreported as some disease pertaining to the digestive system (group 12). The non-medical investigators seldom report non-dysentery cases as dysentery cases. Consequently, the incidence rate for dysentery obtained by non-medical investigators will tend to be biased towards a lower value than the true one. On the other hand, in the case of medical investigators the rate obtained may be regarded as almost nearly the true value mainly due to some of the non-dysentery cases being returned as dysentery cases.

5.16. The group 'other infective and parasitic diseases' is evidently a heterogeneous one. This includes all infectious and parasitic diseases other than pulmonary tuberculosis, malaria and dysentery. Naturally, one would expect much less discrepancy in this particular group. Curiously enough, the non-medical reports are totally discrepant from the corresponding entries in the hospital register. However, a number of diseases belonging to other groups have been reported by the non-medical investigators as diseases belonging to this group. The performance of the medical investigating team in respect of reporting diseases of this category seems to be somewhat satisfactory. About 55% of the cases are reported correctly and only one case belonging to another group has been brought into this category. It is likely, therefore, that the estimate obtained from the medical team will be erring on the lower side. The misclassification in this group usually arises due to neuritis cases being classified as rheumatism, enteric fever as bronchitis or other respiratory diseases. Probably, hospital staff have a tendency to enter any disease of unknown etiology as neuritis which on further examination is reported by the medical investigators as rheumatism or some other specific disease.

5.17. The performance of the medical team in respect of reporting diseases belonging to group 5 (allergic, endocrine, metabolic and nutritional diseases) seems to be fairly satisfactory. But the classification based on the returns of the non-medical investigators seems to be far from satisfactory, only about 16 per cent of the total number of cases allotted to them having been correctly classified. Nevertheless, the rate based on their reports is very nearly equal to the one based on hospital entries for the reason that a number of cases belonging to other groups have been brought into this category due to their peculiar nature of reporting.

5.18. Diseases of blood and blood-forming organs like anaemia are most often misreported as diseases of the digestive or genito-urinary system by the medical and non-medical investigators. There is less tendency on the part of the non-medical investigators to classify diseases belonging to other groups as diseases of this group with the result that the overall estimate of diseases of blood and blood-forming organs will still remain grossly underestimated.

5.19. Diseases belonging to group 7 (mental, psychoneurotic, nervous system etc.,) tend to be misclassified as diseases of the digestive or genito-urinary system or rheumatism though the rates are kept up by the inclusion of diseases of other categories in this group.

5.20. So far as the diseases of the circulatory system are concerned both the types of investigators show equal extent of disagreement with the hospital entries.

5.21. When diseases of the respiratory system (groups 9, 10 and 11) are considered individually the classification based on the reports of the medical investigators seems to be



only moderately good. A closer examination will reveal that the misclassifications are mostly confined within the three groups themselves so that if these three groups are combined into a single group representing all respiratory diseases, the reporting of the medical investigators tends to be more satisfactory. But in the case of non-medical investigators, the misclassifications appreciably extend beyond the three groups and the resulting rate obtained from their reports is likely to be an under-estimate as the losses to these groups are usually higher than the gains.

5.22. The diseases of the digestive system are very often misreported by the medical investigators as dysentery, or diseases of the genito-urinary system. However, this group gains at the expense of diseases like dysentery, anaemia and bronchitis. Hence, the overall rate obtained for this group seems to be very close to that obtained from the hospital register. There is a greater degree of agreement observed in the returns of the non-medical team. But the rate obtained from these reports appears to be grossly exaggerated due to the inclusion of a large number of cases of dysentery and respiratory diseases and to a lesser extent cases belonging to other disease groups in this category.

5.23. The number of cases of diseases of the genito-urinary system obtained from the hospital records being very few, the nature of misclassification cannot be assessed. However, a number of other diseases have been reported as diseases of this category both by the medical and non-medical teams, indicating that the rates obtained on the basis of these reports are likely to be grossly exaggerated.

5.24. In respect of diseases of bones and organs of movement, the extent of disagreement between the hospital diagnoses and the investigators' reports seems to be moderate for both the sets of investigators. But the rates obtained on the basis of their reports for this category of diseases are likely to be exaggerated appreciably due to the inclusion within this group of cases belonging to other groups.

5.25. In the preceding paragraphs it was shown that the extent of misclassification resulting from disagreement between the investigators' reports and the corresponding hospital entries was slightly less in the case of medical investigators than that of non-medical investigators. One may argue that the advantage of using a medical investigator is considerably lowered due to the inclusion of non-prevailing cases. Probably, if the cases were prevailing a more thorough examination of the cases could have been made by the medical investigators which would result in an appreciable improvement in the quality of their reports. If this were so, we may reasonably assume, that for the prevailing cases there should be an additional degree of agreement between the medical investigators' reports and the corresponding hospital diagnoses. For studying this aspect only the cases of diseases prevailing at the time of investigation have been considered and the evaluation of the agreement for the medical and non-medical teams have been made in Table 5.6.

5.26. The results show that an overall disagreement of 55.9 per cent have been recorded for the medical investigators as against 57.3 per cent for them when both the prevailing and non-prevailing cases were considered. When an assessment of the extent of disagreement is done disease-wise for the medical investigators, we find that the divergence is more or less the same as when the non-prevailing cases were also included for all the disease groups except for diseases of the circulatory system, influenza, bronchitis and diseases of the digestive system. In the case of influenza, it has turned out to be worse and in the case of diseases of the circulatory and digestive systems and bronchitis it has turned out to be better.



# A PILOT HEALTH SURVEY IN WEST BENGAL—1955

TABLE 5.6. PERCENTAGE MISCLASSIFICATION AMONG THE MEDICAL AND NONMEDICAL INVESTIGATORS IN DIFFERENT DISEASE CATEGORIES WHEN ONLY PREVAILING CASES WERE CONSIDERED

disease	medical investigator	non-medical investigator
(1)	(2)	(3)
1. group I—tuberculosis (pulmonary)	100.0	50.0
2. group II—malaria	80.0	100.0
3. group III—dysentery	72.7	91.7
4. group IV—other infective and parasitic diseases	37.5	100.0
5. group V—allergic, endocrine system, metabolic and nutritional diseases	22.2	77.8
6. group VI—diseases of blood and blood-forming organs	100.0	83.3
7. group VII—mental, psychoneurotic and personality disorders and diseases of nervous system and sense organs	83.3	88.9
8. group VIII—diseases of circulatory system	33.3	33.3
9. group IX—influenza	100.0	100.0
10. group X—bronchitis	33.3	66.7
11. group XI—other respiratory diseases	100.0	88.9
12. group XII—diseases of digestive system	33.3	22.6
13. group XIII—diseases of genito-urinary system	—	—
14. group XIV—diseases of bones and organs of movement	33.3	0.0
15. group XV—other diseases	—	—
16. total	55.9	64.0

5.27. The analysis in respect of the non-medical investigators revealed that the divergence compared fairly well with that observed when both the prevailing and non-prevailing cases were included in the analysis. However, for diseases like malaria and influenza the degree of disagreement was higher while for diseases of the circulatory system and of the bones and organs of movement the performance of the non-medical team was better when only prevailing cases were considered.

5.28. While it is to be admitted that the performance of the medical investigators seems to be slightly superior to that of the non-medical investigators, there is no indication that the degree of precision in reporting diseases will be enhanced if only the prevailing cases were investigated by the usual questionnaire method without taking recourse to other aids such as physical examination and laboratory tests.

5.29. So far the analysis of the data were based on 15 diseases or groups of diseases. A further condensation of groups though not desirable from the point of view of detail, is expected to result in a closer agreement between the returns of the investigators and the hospital diagnoses. In order to assess the improvement in reporting effected by grouping



the diseases on a still wider basis, the above 15 categories were reclassified into 9 broader categories of diseases. This type of grouping no doubt introduces a greater degree of heterogeneity within the groups. However, the results of the analysis based on such a broad classification will indicate the level to which the extent of disagreement could be brought down.

5.30. Table 5.7 gives the percentage of disagreement observed in the returns of the medical and non-medical investigators when they were compared with the corresponding hospital diagnoses.

TABLE 5.7: PERCENTAGE MISCLASSIFICATION AMONG THE MEDICAL AND NON-MEDICAL INVESTIGATORS IN DIFFERENT DISEASE GROUPS

disease group	medical investigator	non-medical investigator
(1)	(2)	(3)
1. group I— <i>infective and parasitic diseases</i>	67.4	64.4
2. group II— <i>allergic, endocrine system, metabolic and nutritional diseases</i>	22.2	81.8
3. group III— <i>diseases of blood and blood-forming organs</i>	87.5	87.5
4. group IV— <i>mental, psychoneurotic and personality disorders, diseases of nervous system and sense organs</i>	75.0	81.8
5. group V— <i>diseases of circulatory system</i>	50.0	50.0
6. group VI— <i>diseases of respiratory system</i>	26.2	47.9
7. group VII— <i>diseases of digestive system</i>	45.5	23.5
8. group VIII— <i>diseases of bones and organs of movement</i>	42.9	28.6
9. group IX— <i>other diseases</i>	66.7	—
10. total	49.0	53.0

5.31. Considering the diseases which have been merged to form broader groups, it could be said that as far as diseases of the respiratory system are concerned the performance of the medical investigators seems to be better than that of the non-medical investigators, while no appreciable difference is observed with respect to infective and parasitic diseases. Also the overall performance of the medical investigators seems to be slightly superior to that of the non-medical investigators. As expected, the adoption of this grouping has brought about a reduction of nearly 10 per cent in the disagreement rate in comparison to the rate observed when a more detailed classification of diseases (Table 5.3) was considered for both the investigating teams.

5.32. *Influence of social status of the informant on the accuracy of reporting the cause of disease.* To examine whether the social status of the informant has any bearing on the quality of reporting of diseases occurring among the members of his household, the distribution of the informants in each of the educational and occupational groups according to the nature of the disease classification has been obtained and presented in Tables 5.8 and 5.9 respectively for both the medical and non-medical investigators. As most of the patients attending the out patients' department of the hospital to which our data relate belong to the lower



# A PILOT HEALTH SURVEY IN WEST BENGAL—1955

social groups the occupational stratification adopted in the analysis cannot be expected to show significant class differentiation. For instance, it was necessary to include a few professionals in the same group consisting of clerical and other low income groups in order to obtain an appreciable number in the 'high' occupational class. The rest being mostly manual labourers living in bustees had to be allocated to the two classes 'medium' and 'low' according to the skill involved in the jobs.

TABLE 5.8. DISTRIBUTION OF INFORMANTS ACCORDING TO EDUCATIONAL STATUS AND NATURE OF DISEASE CLASSIFICATION

educational status	medical			non-medical		
	complete agreement	no agreement	total	complete agreement	no agreement	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. illiterate	4 (44.4)	5 (55.6)	9 (100.0)	10 (29.4)	24 (70.6)	34 (100.0)
2. literate with no knowledge of English	22 (42.3)	30 (57.7)	52 (100.0)	27 (35.5)	49 (64.5)	76 (100.0)
3. literate with knowledge of English	41 (42.7)	55 (57.3)	96 (100.0)	25 (43.1)	33 (56.9)	58 (100.0)
4. total	67 (42.7)	90 (57.3)	157 (100.0)	62 (36.9)	106 (63.1)	168 (100.0)

TABLE 5.9. DISTRIBUTION OF INFORMANTS ACCORDING TO OCCUPATIONAL STATUS AND NATURE OF DISEASE CLASSIFICATION

occupational status	medical			non-medical		
	complete agreement	no agreement	total	complete agreement	no agreement	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. low	17 (38.6)	27 (61.4)	44 (100.0)	19 (27.5)	50 (72.5)	69 (100.0)
2. medium	25 (41.7)	35 (58.3)	60 (100.0)	23 (37.7)	38 (62.3)	61 (100.0)
3. high	25 (47.2)	28 (52.8)	53 (100.0)	20 (52.6)	18 (47.4)	38 (100.0)
4. total	67 (42.7)	90 (57.3)	157 (100.0)	62 (36.9)	106 (63.1)	168 (100.0)

5.33. As far as the medical investigators are concerned, there is no evidence of any association between the accuracy of their returns and the educational or occupational status of the informants. This suggests that their method of interrogation was practically independent of what the informant stated about the nature of the disease as he understood from the attending physician. However, it can be seen from the above two tables that



the returns of the non-medical investigators were to some extent influenced by the social status of the respondent indicating thereby, that the persons belonging to the higher social class are frequently appraised of the nature of the disease by the attending physicians.

5.34. Two important factors emerge from the results of the Validity Survey discussed in the preceding paragraphs. First, the inaccuracy arising out of misreporting of diseases is considerable in an investigation of this type and second, the accuracy of the disease reporting is not appreciably enhanced by the employment of medical investigators for this purpose. Further, the results shown in Tables 5.4 and 5.5 indicate the directions in which misreporting of diseases takes place which may be fruitfully applied in interpreting morbidity rates in respect of different disease groups. But it is necessary to emphasize in this context that the above Validity Survey included within its scope only such cases of diseases which were attended by an hospital. If a health survey is carried out in a population, it may be observed that a substantial number of diseases occurring in the population do not receive any medical treatment. It is only reasonable to expect a greater degree of inaccuracy in the reporting of such diseases which will only tend to make the situation worse. Moreover, there are no means of checking the validity of non-attended cases except by a prevalence survey by trained medical personnel. But such a survey will have to be necessarily restricted to chronic diseases because they alone can be expected to prevail in appreciable numbers at the time of investigation.

5.35. Though the results of the Validity Survey discussed above give only a partial picture of the inaccuracies in the morbidity returns, it is assumed that they may be of considerable value in the interpretation of the morbidity rates estimated from the data relating to the West Bengal Health Survey. A brief discussion of these rates based on the results of the Validity Survey is attempted in the following paragraphs.

5.36. The West Bengal Health Survey showed a total of 604 cases of illnesses among the members of 1172 rural households and 351 cases of illnesses among the members of 566 urban households during the three-month period of observation beginning in March and ending in May, 1955. As could be expected, some of these illnesses had their onset earlier to the first reference period and some continued to prevail at the close of the last (fourth) reference period. In the former case, if the illnesses were chronic in nature, only the approximate month of onset rather than the exact date was noted in column 6 of block 7 of the schedule. As regards the latter, not even an approximate estimation of duration of illnesses could be availed

5.37. The allocation of illnesses into chronic and acute was done on the basis of their duration. All illnesses whose duration exceeded three months were classified as chronic and those illnesses which prevailed for periods shorter than 3 months were classified as acute. The classification of such illnesses which were prevailing during the last visit and whose duration fell short of 3 months till that date was carried out with the help of medical experts.

5.38. Since the exact time of onset and recovery of acute diseases are more or less abrupt and recognisable, it is customary to define the morbidity of the population in respect of such diseases by the incidence rate which implied the frequency of new cases arising in a given interval of time among 1000 population. This gives a more or less dynamic picture of morbidity and as such the preventive aspects are fully brought to light.



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

5.39. In the case of chronic diseases, however, no abrupt time of onset is recognisable and sometimes the diseases are clinically evident only at an advanced stage. It is rather impossible in such cases to calculate incidence rates and the best that can be done under the circumstances, is to define morbidity in terms of their prevalence. The prevalence rate is defined as the number of cases among 1000 population at a given instant of time. The prevalence rate is a useful measure of the extent of chronic diseases in a population prevailing at a given time regardless of the date of onset of the diseases. This measure, no doubt, gives only a cross-sectional picture of the morbidity of the population and it is very much influenced by the curative aspects such as effectiveness to reduce their duration and the stage at which they are diagnosed. The incidence and prevalence rates for acute and chronic diseases are presented in Tables 5.10 and 5.11 respectively. The reliability of the rates given in these tables can, however, be assessed by a comparison of similar rates obtained from two independent sub-samples shown in Tables 01.2 and 01.3 in Appendix 1

TABLE 5.10. INCIDENCE RATES FOR ACUTE DISEASES CLASSIFIED ACCORDING TO DISEASE GROUPS IN WEST BENGAL

disease group	incidence rate per 1000 population in a year	
	rural	urban
(1)	(2)	(3)
1. group I—malaria	46.28	14.19
2. group II—dysentery	26.54	36.26
3. group III—diarrhoea, enteritis and other diseases of the digestive system	40.15	74.10
4. group IV—other infective diseases of intestinal tract e.g., typhoid, cholera, diseases due to helminths, etc.	10.21	23.65
5. group V—measles, mumps, small pox, chicken pox	25.86	22.07
6. group VI—common cold, tonsilitis, influenza, fever, pneumonia, bronchitis and other respiratory diseases	140.87	179.72
7. group VII—eye, ear, boil and abscess, cellulitis and dental diseases	25.18	36.26
8. group VIII—others (e.g. anaemias, v.d., vascular lesions affecting central nervous system, rheumatic fever, appendicitis, congenital malformations, accidents, etc.)	12.93	36.26
9. total	328.02	422.51



TABLE 5.11. PREVALENCE RATES FOR CHRONIC DISEASES CLASSIFIED ACCORDING TO DISEASE GROUPS IN WEST BENGAL

disease group (1)	prevalence rate per 1000 population	
	rural (2)	urban (3)
1. group I—tuberculosis (pulmonary)	1.68	3.77
2. group II—diseases of the circulatory and nervous systems viz., arteriosclerotic and degenerative heart diseases, hypertension, diseases of veins, rheumatic fever, psychoneurosis, diseases of nerves	3.69	2.52
3. group III—diseases of the eye, ear, skin, bones and joints	4.02	5.03
4. group IV—diseases of the stomach and duodenum except cancer	2.68	5.87
5. group V—asthma	3.52	4.61
6. group VI—diseases of the genital organs, fistula	2.85	4.61
7. group VII—others, (e.g., v.d., cancer, diabetes, avitaminosis, nephritis, congenital and functional diseases, etc.)	2.01	8.39
8. total	20.45	34.80

5.40. At the outset of the analysis it was our intention to strictly adhere to the International Statistical Classification of Diseases and Injuries (List C—Special List of 50 causes for tabulation of morbidity—W.H.O., 1948). Subsequently, it was found from the nature of the data collected that even such an abridged list was too detailed for obtaining any reliable morbidity rates. The classification of diseases in the above analysis had, therefore, to be considerably condensed without appreciably damaging the essential features of the prevailing morbidity pattern in West Bengal.

5.41. The most revealing feature of the above tables is that the diseases, whether acute or chronic, occur more often amongst the urban than in the rural residents, the only exceptions being malaria and diseases of the circulatory and nervous systems. Common cold, influenza and other diseases of the respiratory system are the most commonly reported diseases during the reference period both in the rural and urban areas.

5.42. The observed difference in the incidence and prevalence of diseases between rural and urban communities need not always be indicative of the real extent of variation. In interpreting results of this type, it is necessary to bear in mind the relative importance of the factors likely to influence morbidity returns. Amongst the factors may be mentioned the age-sex composition, the level of health consciousness and the organisation of medical care services in the communities to be compared. Such being the case, any hasty conclusion as to the relative healthiness of two areas without properly weighing these factors may be misleading. In addition to these factors the morbidity rates for specific groups of diseases are appreciably affected by errors due to misclassification as shown by the results of the Validity Survey.



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

5.43. The age-sex composition of the rural and urban samples are shown in Table 5.12.

TABLE 5.12. AGE-SEX DISTRIBUTION OF THE SURVEYED POPULATION

age (in years)		rural		urban	
1. less than 1	persons	152	(2.55) <sup>1</sup>	53	(2.22)
	males	78	(1.31)	27	(1.13)
	females	74	(1.24)	26	(1.09)
2. 1—4	persons	770	(12.91)	273	(11.45)
	males	396	(6.64)	143	(6.00)
	females	374	(6.27)	130	(5.45)
3. 5—14	persons	1478	(24.77)	522	(21.89)
	males	808	(13.54)	278	(11.66)
	females	670	(11.23)	244	(10.23)
4. 15 and above	persons	3566	(59.77)	1537	(64.44)
	males	1804	(30.24)	888	(37.23)
	females	1762	(29.53)	649	(27.21)
5. all age-groups	persons	5966	(100.00)	2385	(100.00)
	males	3086	(51.73)	1336	(56.02)
	females	2880	(48.27)	1049	(43.98)

<sup>1</sup> Figures in parentheses are percentages

5.44. From the above table it appears that the rural and urban samples had more or less similar age-sex composition from which it follows that the urban-rural differentials in morbidity could not be ascribable to the difference in the age-sex composition of their populations.

5.45. A further breakdown of the morbidity rates by age, sex, living conditions, educational and occupational status would, indeed, be helpful in preventive public health work. This was not attempted here as the scope of the survey did not allow such a detailed study.

5.46. Lal and Seal (loc. cit.) have given morbidity rates for certain principal chronic and acute diseases. It may be of interest to make broad comparisons between the morbidity rates estimated from the data of West Bengal Health Survey, 1955, and the Singur Health Survey, 1944. It is well known that some of the acute diseases have a distinct seasonal pattern. It is, therefore, necessary to allow for this seasonal influence while estimating the total number of cases that may be expected during the whole year. No such adjustment for seasonal fluctuations need be made in respect of the Singur Health Survey data, because the information collected relates to one year.

5.47. The morbidity rates of certain important diseases estimated from the rural data collected in this survey have been compared with the corresponding rates obtained from the Singur Health Survey in Table 5.13. To make the comparison strictly valid for such diseases as have a seasonal pattern appropriate adjustments have been made. It could be observed that there is a fair degree of agreement between the two sets of figures. The incidence rate for malaria estimated from the present survey even after accounting



for seasonal influence, is strikingly low in comparison to the one obtained by Lal and Seal for Singur. It was stated earlier while discussing the Validity Survey results that malaria showed a tendency to be misreported as other diseases and that diseases other than malaria were less likely to be returned as malaria. The above tendency was observed to be operating to the same extent among the medical and non-medical investigating teams. It is, therefore, reasonable to assume that the difference in the types of investigators employed in the two surveys could not have resulted in a divergence of the magnitude shown in Table 5.13. Hence, it may be reasonably assumed that the difference between the two malarial rates observed is a real one. This is only natural because Singur during the forties was a highly malarial place, though today malaria has been practically controlled there. Moreover, there is a gap of about a decade between the two surveys during which time a reduction in malaria incidence in West Bengal could have taken place due to better health measures. As regards the incidence of measles, the Singur rate appears to be higher than the rate in this survey. The higher rate for measles observed in the Singur population might have been probably due to its high density of population and its proximity to such congested areas as Howrah and Calcutta.

5.48. It has been pointed out earlier that the performance of the medical investigators was more satisfactory than that of the non-medical investigators in respect of respiratory diseases considered as a whole. The rate obtained from the reports of the latter was usually an under-estimate as diseases belonging to this group were more often returned as diseases belonging to other groups. It is, therefore, not unlikely that the rate estimated from the data relating to the West Bengal Health Survey for pneumonia and influenza falls short of the corresponding rate estimated from the Singur Health Survey data as the latter was based on medical investigators' reports. However, the difference in the rates shown in Table 5.13 is not statistically significant.

TABLE 5.13. COMPARISON OF THE RESULTS OF THE WEST BENGAL HEALTH SURVEY (RURAL) AND THE SINGUR HEALTH SURVEY

disease	annual morbidity rate per 1000 population		
	Singur Health Survey, 1944	West Bengal Health Survey, 1955	
		before adjustment for seasonal pattern	after adjustment for seasonal pattern
(1)	(2)	(3)	(4)
<i>acute disease</i>			
1. malaria	260	46	154
2. dysentery and diarrhoea	38	56	69
3. measles	42	20	18
4. pneumonia and influenza	12	9	7
<i>chronic disease</i>			
5. tuberculosis (pulmonary)	1.09		1.68
6. asthma	2.82		3.52
7. diseases of the circulatory and nervous system	3.44		3.69



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

5.49. With regard to certain chronic diseases also it was noticed that the prevalence rates estimated from the West Bengal Health Survey data were in close correspondence with those estimated from the data of the Singur Health Survey.

5.50. As before, it is necessary to interpret the prevalence rates obtained from the two surveys in the light of the results of the Validity Survey. It was found that the medical investigators misreported all the tuberculosis cases as belonging to some other diseases. The non-medical investigators also misreported a substantial number of tuberculosis (pulmonary) cases. But they exhibited a tendency to include a number of cases of other diseases in the tuberculosis (pulmonary) group leading to an exaggerated prevalence rate for this group. Whatever may be the direction in which misclassification of tuberculosis (pulmonary) cases tend to occur, it seems that the only method of assessing accurately the prevalence of pulmonary tuberculosis is by complete physical examination of the population surveyed.

5.51. In respect of allergic diseases like asthma, etc., the medical investigators' performance was found to be far superior to that of the non-medical investigators. As regards diseases of the nervous and circulatory systems the extent of agreement seemed to be almost the same for both the groups of investigators. But the rates based on the non-medical investigators are likely to be exaggerated on account of including in each of the above groups, diseases belonging to other groups.

5.52. In diseases of a chronic nature such as T.B. where there is no abrupt onset of a diseased condition in the affected individuals the estimated morbidity rates should be taken as corresponding to clinically diagnosed diseases or those causing severe disability. Moreover, there are other reasons which tend to under-estimate the morbidity rates of such diseases. For instance, there is a certain amount of time-lag between the actual onset of a disease and the time when medical diagnosis is sought. The degree of disability or discomfort arising out of an affliction and the level of health consciousness of the subjects are among the important factors which largely determine the stage at which the disease is subjected to a medical diagnosis. It is, therefore, inevitable that some of the cases go unaccounted due to the operation of these and similar factors. In some acute diseases the subjects may fail to recognise the symptoms manifested by these diseases due to their ignorance and low health consciousness. Hence, interpretation of morbidity rates have to be based on a proper appreciation of the factors involved. In the report of the Sickness Survey conducted in U.K. by the Ministry of Health (1946), it was observed that out of a sample of about 7,000 population, 5,518 or 79% suffered from one or more illnesses or injuries during a three-month period. Pearse and Crocker (1944) in their study 'The Pekham Experiment— a study of the living structure of society' have also arrived at similar results. They estimated that only about 10 per cent of the population on which an health overhaul was done was without any sign of disorder and the remaining 90 per cent were either in disease or in whom disorder was associated with a sense of well-being. As against this, the estimated number of cases of illnesses or injuries during the three-month period in rural West Bengal was 604 out of a sample of 5,966 persons i. e., 10 per cent and in urban West Bengal the corresponding number was 351 out of a sample of 2,385 persons i.e., 15 per cent.

5.53. The contrast between the estimates of West Bengal and U.K. is indeed striking. The results suggest that the people of U.K. are less healthy than those of West



Bengal which contradicts the prevailing notion about the relative levels of health of these two populations. These findings have to be interpreted in the light of the health consciousness of the subjects which is essentially a concomitant of their levels of living. As there is no well defined line of demarcation between the state of health and that of disease of an individual it is likely that the morbidity returns obtained from an investigation of this type are influenced appreciably by the level of health consciousness of the community. In our country where the degree of health consciousness is known to be low, there is a natural tendency to overlook minor ailments and report only such conditions which cause pain, discomfort, or disability to the subjects. Hence, a substantial number of illnesses might not have been reported at all. As the morbidity data collected by means of interrogation of the individuals are affected by a considerable degree of subjectivity, the only means of assessing the extent of morbidity seems to be a prevalence survey carried out on the basis of a complete physical examination supplemented by laboratory tests, if necessary.

## CHAPTER 6

### DISABILITY

6.1. In the preceding paragraphs, discussion was chiefly confined to the frequency of incidence and prevalence of diseases among the rural and urban populations of West Bengal and their classification by causes. In what follows, an attempt is made to describe briefly the question of disability arising out of these diseases and their social consequences.

6.2. Though it is desirable to split the duration of disabling illnesses into days of disability and non-disability, it was not possible to do so with the data in hand. Hence, what is referred to as days of disability hereinafterwards is actually the duration of disabling illnesses.

6.3. As stated earlier, illnesses, both chronic and acute, were divided into three classes according to the nature of disability caused by them, namely, (i) non-disabling (ii) disabling but not causing confinement to bed or hospital and (iii) causing confinement to bed or hospital.

6.4. The illnesses which did not cause confinement to bed or hospital were classified as disabling (case (ii) above), if the illnesses led to either stoppage of usual work or availing of medical care or special diet. Otherwise, they were classified as non-disabling. Consequently, the concept of disability is somewhat less objective for children and for aged persons who generally do not have any particular assignment of work in or outside the household. Hence, for a critical evaluation of disability and its consequences, the discussion is limited to persons in the age-group 15-59 years. Moreover, since most of the persons belonging to this age-group are either in the labour force or engaged in domestic duties, the disability arising in this segment of the population may have inevitable economic and social consequences.

6.5. The total number of disabling illnesses and their proportion to total illness occurring to the rural and urban populations in the age-group 15-59 years are shown in



# A PILOT HEALTH SURVEY IN WEST BENGAL—1955

Table 6.1. The reliability of these figures can be assessed by a comparison of the sub-sample estimates shown in Table 01.4 of Appendix 1.

TABLE 6.1. ILLNESSES OCCURRING DURING THE REFERENCE PERIOD  
CLASSIFIED INTO TYPE OF DISABILITY IN THE AGE GROUP  
15-59 YEARS

sector	non-disabling illness	disabling illness	total illnesses	percentage of non-disabling illnesses
(1)	(2)	(3)	(4)	(5)
1. rural	115	217	332	34.64
2. urban	49	135	184	26.63
3. total	164	352	516	31.78

6.6. The proportions of non-disabling illnesses to total illnesses are about 35 per cent and 27 per cent for the rural and urban populations respectively. In the survey of sickness of the population of U.K. (*loc. cit.*) it was observed that out of 4667 cases of independent illnesses occurring to the adult population in the sample, 4237 or 91 per cent were of a non-disabling nature or had duration of disability for less than a day. In a morbidity study carried out in the Eastern Health District of Baltimore during 1938-43, it was observed that 53 per cent of total cases of all ages were non-disabling. A comparison of the West Bengal Survey results with those pertaining to the U.K. or Baltimore clearly indicates that a number of non-disabling illnesses have not been reported in the West Bengal Survey, a substantial fraction of which, it may not be unreasonable to attribute to the low level of health-consciousness of the people. If by some method this unknown number can be estimated and added to the already reported non-disabling cases, one could have had an approximate estimate of the number of people in an indifferent state of health, who could not pull their full weight in the economic activities in which they are engaged.

6.7. The results presented in Table 6.1 need not necessarily reflect the real extent of the social and economic implications of disability to the community. A better measure of this may be the duration of disability due to various causes and their frequency of occurrence indicating the extent of human wastage which otherwise could have been utilised for productive purposes. For this purpose, the duration of disability due to each kind of illness falling strictly within the reference period was cumulated over the four reference periods and inflated four times to yield annual estimate of number of days lost due to disability arising from each type of disease. No attempt was, however, made to adjust for the seasonal peculiarity of the survey period. In Table 6.2 are shown the total days of disability in a year for the age-group 15-59 years in the surveyed population. Similar results are given for the two sub-samples in Table 01.5 in Appendix 1.

6.8. Of the acute diseases, malaria, dysentery, diarrhoea, enteritis and other diseases of the digestive system, diseases of the respiratory system and boil, abscess, cellulitis etc., are the principal diseases causing disability in both the rural and urban areas. As could be expected malaria accounts for a higher annual rate of disability in the rural areas than in the urban areas. Similarly, diseases of the digestive system (other than diarrhoea and enteritis) and boil, abscess, cellulitis etc., are associated with higher annual



TABLE 6.2. TOTAL DISABILITY DAYS IN A YEAR AND DISABILITY DAYS PER PERSON IN A YEAR IN THE SURVEYED POPULATION AGED 15-59 YEARS

disability due to	rural		urban	
	total disability days in a year	disability days per person in a year	total disability days in a year	disability days per person in a year
(1)	(2)	(3)	(4)	(5)
<i>acute diseases :</i>				
(i) malaria	1,108	0.34	372	0.26
(ii) dysentery	572	0.17	293	0.20
(iii) diarrhoea and enteritis	246	0.07	409	0.28
(iv) other acute diseases of digestive system	738	0.22	139	0.10
(v) acute diseases of respiratory system including fever	2,448	0.75	1,267	0.88
(vi) boil, abscess, cellulitis and other skin infections	3,183	0.97	692	0.48
(vii) other acute diseases	1,969	0.60	1,669	1.16
all acute diseases	10,264	3.12	4,841	3.36
all chronic diseases	14,965	4.55	14,600	10.10
all diseases	25,229	7.67	19,441	13.46

disability rates amongst rural persons. On the other hand, the rate of disability due to diarrhoea and enteritis is higher amongst the urban population. The rate of disability due to dysentery, however, does not show any sharp differential between the rural and urban groups.

6.9. The overall annual rate of disability in terms of days lost due to either acute or chronic diseases is more for the urban sector than for the rural sector. Comparing the estimate arrived at for the urban sector with those obtained for the Sickness Survey in U.K. (16.8 days per adult annually) and the morbidity survey in the Eastern Health District of Baltimore (15.9 days per person annually) the West Bengal figure seems to be an under-estimate. The rural-urban difference in the rates of disability due to acute diseases is not so pronounced as in the case of chronic diseases. For example, an urban adult on an average loses nearly twice the number of days on account of disability arising out of chronic diseases as compared with a rural adult. Thus, the chronic diseases which are relatively infrequent in terms of cases are particularly important to the community as well as to the individual patient in terms of days lost from usual activities especially in urban areas. As most of the reported illnesses, both acute and chronic, are preventable, it is possible with the successful application of public health measures to raise the level of health of the population. This will in turn increase its potential efficiency leading ultimately to a higher level of living.



## CHAPTER 7

## MEDICAL AND MATERNITY CARE

7.1. *Medical care.* The frequency with which diseases occur and their duration and the nature of disability in a community are no doubt of great value to the public health administrator. But they describe only one side of the health picture of the population. On the other hand, the amount of medical care available to the population roughly indicates whether such facilities are sufficient to cope with the morbidity situation. Secondly, the extent to which they are utilized by the afflicted individuals suggests how the process of recovery is affected. Thirdly, knowledge as to who are the actual beneficiaries of the existing medical set-up will considerably help in planning the distribution of medical benefits to the population.

7.2. Since the achievement of independence of India, the importance and urgency of providing adequate medical care in its curative and preventive aspects are increasingly realised. The Health Survey and Development Committee (*loc. cit*) rightly points out that 'a nation's health is perhaps the most potent single factor in determining the character and extent of its development and progress and any expenditure of money and effort on improving the national health is a gilt-edged investment yielding immediate and steady returns in increased productive capacity. . . . The provision of adequate health protection to all covering both its curative and preventive aspects, irrespective of their ability to pay for it, . . . are all facets of a single problem and call for urgent attention.'

7.3. The situation as it exists today is far from satisfactory. The high incidence of preventable diseases and the heavy toll of life taken by these diseases, the abnormal infant and maternal mortality, the widespread existence of malnutrition and under-nutrition, deplorable housing conditions, and grossly inadequate preventive and curative health services are important features of the present health picture of the population. A comparison of the existing medical facilities in our country with those available in the more progressive countries like the U.K. or the U.S.A. (Table 7.1) will reveal how inadequate the available facilities are.

TABLE 7.1. MEDICAL PERSONNEL AND HOSPITAL FACILITIES IN INDIA, U.S.A. AND U.K.

country	year	inhabitants per			
		physician	midwife	pharmacist	hospital bed
(1)	(2)	(3)	(4)	(5)	(6)
U.S.A.	1953	750	400 <sup>1</sup>	1,600	100
U.K.	1951	1,150	4,550	3,500	90
India	1952	5,700	23,000	13,700	2,450

<sup>1</sup> Refers to graduate nurses in 1954.

7.4. It is clear from the above table that India lags far behind the U.S.A. and the U.K. as regards the availability of medical personnel and hospital beds. The worst sufferers



are the rural population of India. They form about 80 per cent of the total population but hardly 30 per cent of the doctors are available to them. The situation is equally bad with respect to hospital beds and dispensaries. Further, as the rural population is widely dispersed with no adequate transport facilities, accessibility to the medical personnel, hospitals or dispensaries is considerably restricted.

7.5. As an auxiliary to the present health survey, an investigation into the availability of medical facilities in rural areas was conducted in 69 villages selected for the survey. The results of the investigation shown in Table 7.2 help to evaluate approximately the extent of medical care available in rural West Bengal at present. A comparison with official figures for all West Bengal will reveal the rural-urban differential in respect of the availability of medical personnel.

TABLE 7.2. THE AVAILABILITY OF REGISTERED MEDICAL PRACTITIONERS IN THE SURVEYED RURAL POPULATION AND IN WEST BENGAL

population	no. of villages in sample	total population of villages	no. of regd. doctors		inhabitants per regd. doctor		
			allopath	all systems	allopath	all systems	all West Bengal, 1951 (allopath) <sup>1</sup>
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. less than 1000	41	19,827	2	2	9,414	9,414	
2. 1000—1999	14	18,707	4	6	4,677	3,118	
3. 2000 and above	14	61,658	19	21	3,245	2,936	
4. total	69	100,192	25	29	4,008	3,455	1,318

<sup>1</sup> Statistical Abstract of West Bengal—1952.

7.6. The usual way of presenting the amount of medical facilities available to the population as in Table 7.1 cannot be considered as appropriate especially in countries like India where there is no national health service catering to the needs of the entire population as in the Western countries. In India where an overwhelming majority of the population cannot pay for medical care and the available free medical institutions run by the government or charitable societies are too inadequate to meet the needs of these people, a description of the facilities in terms of population cannot give a true picture of medical care at the disposal of the really needy. As the medical man-power like the rest of the population is attracted by areas of greatest economic and social advantage, the poorer sections of the population who are the vulnerable groups from the point of view of morbidity, might not be able to avail of any sort of medical treatment for pecuniary reasons. Therefore, to obtain a true picture of medical care pattern in a community, it is essential to have besides an assessment of such facilities available to the community, an assessment of the extent of facilities actually availed by the community.

7.7. During the course of the three-month observational period information on 482 illnesses of an acute nature and 122 illnesses of a chronic nature were gathered from the canvassed rural households. Similarly, during the same period, data on 268 acute illnesses and 83 chronic illnesses were collected from the canvassed urban households. The



# A PILOT HEALTH SURVEY IN WEST BENGAL—1955

proportion of illnesses receiving medical treatment and type of such treatment are given in Table 7.3. The reliability of these estimates can, however, be assessed by comparing similar estimates obtained from the two sub-samples shown in Table 01.6 of Appendix 1.

TABLE 7.3. PERCENTAGE DISTRIBUTION OF DISEASES OR INJURIES ACCORDING TO TYPE OF TREATMENT RECEIVED

type of treatment	rural		urban	
	acute	chronic	acute	chronic
(1)	(2)	(3)	(4)	(5)
1. allopath	39.42	46.72	42.54	79.52
2. homeopath	16.18	11.48	23.88	12.05
3. ayurved or unani	6.22	9.84	3.36	7.23
4. quack or no treatment	39.83	40.98	33.58	25.30
5. total	101.65 <sup>1</sup> (482) <sup>2</sup>	109.02 (122)	103.36 (268)	124.10 (83)

<sup>1</sup> Percentages will add up to more than 100, as some cases received more than one type of treatment.

<sup>2</sup> Figures in parentheses are the numbers of cases reported during reference period.

7.8. It is found that about 41 and 51 per cent of the cases in the rural and urban areas respectively availed allopathic treatment whereas only about 7 and 4 per cent of the cases took recourse to the Indian system of medicine. About 15 per cent of the rural cases and 21 per cent of the urban cases availed homeopathic treatment. All these suggest that allopathic system of treatment is more commonly availed by the population even in the rural areas. Between the homeopathic and Indian system of medicine, the former is the more popular one judging from the proportion of cases treated. Of course, the popularity of any system of treatment is the combined effect of efficiency, cost and availability.

7.9. Another fact brought out clearly by the above table is regarding the proportion of illnesses medically attended. It is found that only about 40 per cent of the rural cases and about 32 per cent of the urban cases did not avail treatment from any recognised medical system. This is really surprising in view of the fact that even in such advanced countries like the U.K. or Canada where health services have reached a high level of development, the proportion of cases not seeking medical care is much higher. For instance, in the sickness survey done in the U.K., (*loc. cit.*) it was observed that about 60 per cent of the cases did not avail of any kind of treatment. In the Canadian Sickness Survey, 1950-51, it was estimated that out of a total of 29,471 complaint periods 21,134 or about 72 per cent received no health care. In contrast to these estimates, the West Bengal Health Survey has shown a very low figure for the proportion of cases, not availing any medical treatment. Considering that in West Bengal as in other parts of India, there is a paucity of medical personnel, hospital beds and other treatment facilities compared to those medically advanced Western countries, it is somewhat difficult to reconcile the observed result. The possible explanation for this, as has been stated earlier, may be found in the tendency to omit minor illnesses or injuries causing little or no disability and for which



perhaps no medical care was sought. If by some means it is possible to estimate such omissions, the proportion not availing any medical care is bound to go up.

7.10. The same situation could be viewed from another angle to ascertain the real extent of medical care availed. It is not unreasonable to assume that the morbidity rate in West Bengal is higher than that of the U.K. or the U.S.A. and that almost all cases medically treated in West Bengal are generally reported. Under the circumstances, the ratio of treated cases to the total population will furnish a better appraisal of the extent of medical care availed. In West Bengal it was found that 650 cases out of a total of 955 cases occurring in a period of 3 months to 8351 persons comprising the rural and urban populations, received some kind of medical attention. In other words, 7.8 per cent of the population could avail of medical care. The corresponding figure as revealed in the Sickness Survey in U.K. (*loc. cit.*) is about 31 per cent and that of the Canadian Sickness Survey, 1950-51 (*loc. cit.*) is about 53 per cent. That is suggestive of the fact that the number actually seeking medical advice is significantly low inspite of indications to the contrary that about two-thirds of the cases are medically attended.

7.11. It would have been useful to analyse the data used in Table 7.3 by further breakdowns for disease groups, occupations etc., for a proper appreciation of the medical care pattern availed by the community. The scope of the available data, however, restricts an analysis of this nature.

7.12. Those who did not avail of any sort of medical treatment during their illness were further asked to state the reason(s) for not doing so. In both the rural and urban groups about 41 per cent of such persons attributed it to sickness being 'not serious'. About 33 per cent of the unattended rural cases stated that 'medical care was too expensive' whereas the comparable figure for the urban group was only 6 per cent. It is natural that the abject poverty of the rural population only tend to make medical care too expensive.

7.13. Another important aspect of medical care is regarding the expenses incurred on medical treatment. Here again, a detailed analysis showing the average expenditure incurred with respect at least to the more commonly occurring diseases will be really useful. But for reasons stated earlier such an analysis is not attempted.

7.14. The cost of medical care in terms of expenditure incurred is higher for an urban case than for a rural case. This is evident from Table 7.4 which gives the expenditure incurred per case during the observational period of three months for different types of treatment. Some idea of the reliability of the estimates can be had by comparing the two sub-sample estimates given in Table 01.7 of Appendix 1.

7.15. The higher cost of medical care in towns and cities may be primarily due, among others, to the superior quality of medical facilities available to the urban population. Also, there is a natural tendency for cases when they become advanced in stage to migrate to urban areas in quest of better treatment. Obviously, a higher expenditure is involved in the treatment of such cases.

7.16. It is probably reasonable to bear in mind while interpreting these figures that they are likely to be exaggerated because the usual tendency is to include not only the actual expenses incurred during the period under review but also expenditure relating to some previous period cleared off during the reference period.



# A PILOT HEALTH SURVEY IN WEST BENGAL—1955

TABLE 7.4. MEDICAL EXPENDITURE (IN RS.) INCURRED PER ILLNESS DURING THE REFERENCE PERIOD ACCORDING TO NATURE OF ILLNESS AND TYPE OF TREATMENT AVAILABLE

type of treatment	rural		urban	
	acute	chronic	acute	chronic
(1)	(2)	(3)	(4)	(5)
1. allopath	8.87 (190) <sup>1</sup>	36.89 (57)	28.96 (114)	75.90 (66)
2. homeopath	4.65 (78)	15.79 (14)	13.45 (64)	45.38 (10)
3. ayurved or unani	5.30 (30)	33.23 (12)	80.11 (9)	37.95 (6)
4. quack or no treatment	1.51 (192)	2.79 (50)	7.28 (90)	7.39 (21)
5. total	5.18 (482) <sup>2</sup>	23.46 (122)	20.66 (268)	70.43 (83)

<sup>1</sup> Figures in parentheses are the numbers of cases on which the estimates are based.

<sup>2</sup> Totals will not tally as some cases received more than one type of treatment.

7.17. *Maternity care.* Considerable attention is being paid in recent years for the promotion and protection of the health of the mother and child. Comprehensive schemes have been launched for the training of maternal and child health personnel like *dhais*, midwives, health visitors, nurses etc., in appreciable numbers in order to raise the existing maternity services to a satisfactory level in a short time. The Second Five Year Plan envisages the establishment of numerous health centres to look after the interests of the mother and child.

7.18. Data have been collected in this survey to assess the extent and type of maternity services availed by the population and the results are briefly summarised in the following paragraphs.

7.19. Those pregnancies terminating during the three-month observational period were referred to as current terminations in block 8 of the schedule. Only for such terminations detailed information regarding maternity care received have been collected. This restriction had to be imposed because such detailed information could not be elicited if the events related to the distant past. However, by the above restriction the sample became extremely inadequate to yield any reliable estimates. For such studies, therefore, a special survey has to be carried out including only those households in which births are known to have occurred.

7.20. Of the 45 births taking place during the period under review, 25 births took place to rural mothers and 20 births to urban mothers. The inadequacy of trained professional assistance available at deliveries, particularly in rural areas, is clearly revealed by Table 7.5 which gives the distribution of deliveries according to agencies attending them.



TABLE 7.5. PERCENTAGE DISTRIBUTION OF CURRENT TERMINATIONS  
ACCORDING TO TYPE OF ATTENDANCE

sector	doctor	midwife or nurse (qualified)	dhai	hospital	relatives and friends	total <sup>1</sup>
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. rural	8.0	0.0	76.0	0.0	24.0	108.0
2. urban	25.0	10.0	25.0	35.0	20.0	115.0

<sup>1</sup> Percentages will add up to more than 100 as some deliveries received more than one type of attendance.

7.21. As may be seen from the above table, *dhais* attended about three-fourth of rural deliveries and one-fourth of urban deliveries. All the rural deliveries were non-institutional, whereas 35 per cent of urban births took place in hospitals. Professional service of doctors, and qualified midwives or nurses were availed only in 8 per cent of rural cases. The corresponding figure for the urban cases was 35 per cent. No professional assistance was called for in 24 per cent of the rural cases and 20 per cent of urban cases. Though these estimates are based on small numbers there is no gainsaying that rural populations have to rely largely on primitive and untrained agencies for this purpose.

7.22. It was observed that as high as 96 per cent of the rural births and 85 per cent of the urban births were delivered within the same district.

7.23. On an average, medical expenses which may include payment for the services of a doctor, hospital, midwife, nurse, or *dhai* or cost of medicine was about Rs. 9 in the case of a rural birth, and Rs. 22 in the case of an urban birth. The higher average cost incurred in towns or cities is natural because the services available there being of a superior nature are more expensive.

7.24. The average periods of confinement and convalescence were about 11 days and 18 days for a rural mother and 6 days and 13 days for an urban mother. The lower periods of confinement and convalescence of an urban mother need not necessarily reflect that she received less effective post-natal care than her rural counterpart, nor does it indicate that urban mothers attained normalcy earlier. The longer periods of confinement and convalescence observed in the case of rural mothers merely reflect the wider prevalence of social taboos among them.

## CHAPTER 8

### RECOMMENDATIONS

8.1. The present study has highlighted certain features of methodological importance in the conduct of a health survey which may be given due consideration while planning similar studies in future.

8.2. In a general health survey the main emphasis obviously is on the collection of information on the frequency of the incidence or prevalence of illnesses (or injuries) and



classification by either individual causes or groups of causes. Hence, to make an accurate assessment of the morbidity pattern, two conditions require to be satisfied. First, completeness of the morbidity returns and second, correctness of the classification by causes. The results of the Validity Survey have indicated the possibility of some illnesses being not recorded at all. For instance, about 9 per cent of the cases who had sought hospital aid for some ailment had been missed by the investigators. If this could be attributed to the failure of memory of the respondent to recall the event, then it could reasonably be expected that in a health survey of the general population in which a substantial bulk of the afflicted individuals go without any sort of medical attendance, illnesses are liable to be missed to a greater extent. If the households are contacted only once, some information is likely to be lost unless the period of reference is of a short duration. This would inevitably lead to insufficient coverage over time. This difficulty can be got over by planning the survey in such a manner as to make it possible to visit each selected household a number of times at reasonably short intervals.

8.3. It is known that the incidence of certain diseases exhibit a well-defined seasonal pattern. It is, therefore, desirable to spread the survey period over a complete year. Such a long period of observation would necessarily entail an inconveniently large number of visits to the same households which might perhaps create practical difficulties. To avoid this it is desirable to divide the year into 4 typical seasons of 3 months each. The total sample of households may also likewise be split up into 4 sub-samples and each sub-sample allotted to each season.

8.4. This study has indicated that the value of a health survey by the usual questionnaire method to assess the extent of morbidity with respect to diseases like pulmonary tuberculosis is of a questionable nature. Reliance, therefore, has to be placed on prevalence surveys making use of diagnostic facilities including laboratory tests for a proper evaluation of the prevalence of such diseases.

8.5. That a good deal of misreporting of diseases occur is evident from the results of the Validity Survey. Hence, in order to make a reasonably correct interpretation of the results of morbidity returns, it is suggested that a validity survey to assess the extent and direction of misclassification of diseases be simultaneously attempted.

8.6. In an investigation of this kind the personal error of the investigators is nonetheless a major factor in determining the reliability of the results. An internal check of the sample which takes account of not only the sampling variance but also the personal error of the investigators is, therefore, necessary to establish the reliability of the final estimates. This is easily provided by dividing the entire sample into a series of interpenetrating sub-samples.

8.7. Socio-economic factors like education and occupation have been found to be useful criteria for stratifying urban populations. However, they have their limitations when applied to rural populations. For a social stratification of rural households, therefore it may be desirable to take into consideration such factors as size of holdings or other suitable economic characteristics closely related to the actual living levels of the households.

8.8. This pilot survey comprises of about 1200 rural households selected from about 3.8 million households in rural West Bengal, i.e., one in every 3,000 households. As far as gross morbidity rates were concerned, this sample was found to be adequate to give fairly



precise estimates. But for a detailed analysis with finer breakdowns, the sample size proved to be inadequate.

8.9. If a nation-wide morbidity survey is attempted and the same sampling fraction as above is maintained, then it would be possible to obtain precise estimates of morbidity rates by finer breakdowns at the all-India level, and at the same time State estimates of gross morbidity rates could be obtained with fair degree of precision.

8.10. The urban sampling procedure requires modification to suit the special features of a general health survey to obtain a more economical design. However, it may not be possible to suggest an adequate sample size for the urban population until some more pilot surveys are conducted.

8.11. General health surveys are not expected to provide adequate data for a detailed study of various aspects relating to maternity unless the sample is made unduly large. For such studies it is desirable that the sampling frame consists of households where births are known to have occurred during a recent specified period.

### ACKNOWLEDGEMENTS

The authors of this report are greatly indebted to Dr. K. N. Mitra, M.D., F.R.C.S., F.R.C.O.G., Professor of Obstetrics and Gynaecology, Calcutta Medical College, for his valuable suggestions and advice in the formulation of the questionnaire. Thanks are also due to Dr. R. N. Moitra, M.B., Medical Officer in charge of the Indian Statistical Institute Medical Welfare Unit, Dr. J. N. Chakrabarty, B.Sc., M.B., and Dr. N. C. Sanyal M.B., for their whole-hearted collaboration in the Validity Survey. Finally, the authors are extremely grateful to the authorities of the R. G. Kar Medical College Hospitals, Calcutta for kindly providing details of the patients attending the O. P. D. for use in the Validity Survey.

### REFERENCES

- COLLINS, S. D., PHILLIPS, F. R., AND OLIVER, D. S. (1950): Specific causes of illness found in monthly canvasses of families. Sample of the Eastern Health District of Baltimore, 1938-43. Pub. Health Rep. 65, 1235-1264.
- (1951): Disabling illness from specific causes among males and females of various ages. Sample of the Eastern Health District of Baltimore, 1938-43. Pub. Health Rep. 66, 1649-1671.
- DOMINION BUREAU OF STATISTICS AND DEPARTMENT OF NATIONAL HEALTH AND WELFARE (1956): Canadian Sickness Survey, 1950-51, 9, Ottawa, Canada.
- DAS GUPTA, A., SOM, R. K., MAJUMDAR, M., AND MITRA, S. N., (1955): Couple fertility: National Sample Survey, No. 7. The Department of Economic Affairs, Ministry of Finance, Government of India.
- GOVERNMENT OF INDIA (1946): Report of the Health Survey and Development Committee, 1-4, Government of India Press, New Delhi.
- LAL, R. B., AND SEAL, S. C. (1949): General Health Survey, Singur Health Centre, 1944, Government of India Press, Calcutta.
- PEARSE, I. H. AND CROCKER, L. H. (1944): *The Peckham Experiment—a study of the living structure of society*, George Allen and Unwin Ltd., London.
- PLANNING COMMISSION (1956): Second Five Year Plan, Government of India.
- ROYAL COMMISSION ON POPULATION (1950): Reports of the Biological and Medical Committee, 4, His Majesty's Stationery Office, London.
- SLATER, P. (1946): Survey of Sickness, October 1943 to December 1945. Ministry of Health, London
- Paper received: June, 1957.*



# A PILOT HEALTH SURVEY IN WEST BENGAL—1955

## APPENDIX 1

### COMPARISON OF TWO INDEPENDENT SUB-SAMPLE ESTIMATES

TABLE 01.1. INFANT MORTALITY RATE PER 1000 LIVE BIRTHS FOR TWO SUB-SAMPLES

sector	sub-sample 1		sub-sample 2		combined	
	no. of live births	infant mortality rate	no. of live births	infant mortality rate	no. of live births	infant mortality rate
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. rural	2,909	173.60	2,630	164.26	5,539	169.16
2. urban	911	155.87	934	116.70	1,845	136.04

TABLE 01.2. INCIDENCE RATES FOR ACUTE DISEASES CLASSIFIED ACCORDING TO DISEASE GROUPS FOR TWO SUB-SAMPLES

disease group	incidence rate per 1000 population in a year					
	rural			urban		
	sub-sample 1	sub-sample 2	combined	sub-sample 1	sub-sample 2	combined
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. group I—malaria	38.81	54.44	46.28	20.94	8.63	14.19
2. group II—dysentery	16.82	31.52	26.54	45.38	28.79	36.26
3. group III—diseases of the digestive system <sup>1</sup>	49.16	30.08	40.15	104.72	48.94	74.10
4. group IV—other infective and parasitic diseases <sup>2</sup>	7.76	12.89	10.21	24.43	23.03	23.65
5. group V—measles, mumps, small pox, chicken pox	31.05	20.06	25.86	34.90	11.52	22.07
6. group VI—respiratory diseases <sup>3</sup>	129.00	153.28	140.87	160.57	195.77	179.72
7. group VII—eye, ear, boil and abscess, cellulitis and dental diseases	19.41	31.52	25.18	62.83	14.40	36.26
8. group VIII—other diseases <sup>4</sup>	11.64	14.33	12.93	27.92	43.19	36.26
9. total	303.65	348.12	328.02	481.69	374.27	422.51

<sup>1</sup> Diarrhoea, enteritis, etc.

<sup>2</sup> Typhoid, cholera, diseases due to helminths etc.

<sup>3</sup> Common cold, influenza, pneumonia, bronchitis, tonsillitis etc., including fever.

<sup>4</sup> Anaemia, v.d., vascular lesions affecting central nervous system, rheumatic fever, congenital malformation accident etc.



TABLE 01.3. PREVALENCE RATES FOR CHRONIC DISEASES CLASSIFIED ACCORD-  
ING TO DISEASE GROUPS FOR TWO SUB-SAMPLES

disease group	prevalence rate per 1000 population					
	rural			urban		
	sub-sample 1	sub-sample 2	combined	sub-sample 1	sub-sample 2	combined
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. [group I—tuberculosis (pulm.)	0.96	2.47	1.68	2.78	4.59	3.77
2. group II—diseases of the circula- tory and nervous systems <sup>1</sup>	3.19	4.24	3.69	3.71	1.53	2.52
3. group III—diseases of the eye, ear, skin, bones and joints.	3.83	4.24	4.02	4.64	5.36	5.03
4. group IV—diseases of the stomach and duodenum except cancer	2.55	2.83	2.68	5.57	6.12	5.87
5. group V—asthma	3.19	3.89	3.52	5.57	3.83	4.61
6. group VI—diseases of the genital organs	3.51	2.12	2.85	4.64	4.59	4.61
7. group VII—other diseases <sup>2</sup>	2.55	1.41	2.01	8.34	8.42	8.39
8. total	19.78	21.20	20.45	35.25	34.44	34.80

<sup>1</sup> Arteriosclerotic and degenerative heart diseases, hypertension, rheumatic fever, diseases of veins, psychoneurosis, diseases of nerves etc.

<sup>2</sup> V.D., cancer, diabetes, avitaminosis, congenital and functional diseases, etc.



TABLE 01.4. ILLNESSES OCCURRING DURING THE REFERENCE PERIOD CLASSIFIED INTO TYPE OF DISABILITY IN THE AGE-GROUP 15-59 YEARS FOR TWO SUB-SAMPLES

sector	sub-sample 1				sub-sample 2				combined			
	non-dis- abling illness	disabling illness	total	percentage of non- disabling illness	non-dis- abling illness	disabling illness	total	percentage of non- disabling illness	non-dis- abling illness	disabling illness	total	percentage of non- disabling illness
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1. rural	63	113	176	35.80	52	104	156	33.33	115	217	332	34.64
2. urban	30	58	88	34.09	19	77	96	19.79	49	135	184	26.63
3. total	93	171	264	35.23	71	181	252	28.17	164	352	516	31.78



TABLE 01.5. TOTAL DISABILITY DAYS IN A YEAR AND DISABILITY DAYS PER PERSON IN A YEAR IN THE SURVEYED POPULATION AGED 15-59 YEARS FOR TWO SUB-SAMPLES

disability due to	rural						urban					
	total disability days in a year			disability days per person in a year			total disability days in a year			disability days per person in a year		
	sub-sample 1	sub-sample 2	combined	sub-sample 1	sub-sample 2	combined	sub-sample 1	sub-sample 2	combined	sub-sample 1	sub-sample 2	combined
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
<i>acute diseases</i>												
1. malaria	535	555	1108	0.32	0.35	0.34	299	73	372	0.45	0.09	0.26
2. dysentery	156	416	572	0.09	0.27	0.17	65	228	293	0.10	0.29	0.20
3. diarrhoea and enteritis	172	74	246	0.10	0.05	0.07	281	128	409	0.43	0.16	0.28
4. other acute diseases of digestive system	615	123	738	0.36	0.08	0.22	76	63	139	0.12	0.08	0.10
5. acute diseases of respiratory system including fever	939	1,509	2,448	0.54	0.97	0.75	416	851	1,267	0.64	1.07	0.88
6. boil, abscess, cellulitis, and other skin infections	777	2,406	3,183	0.45	1.54	0.97	179	413	692	0.43	0.52	0.48
7. other acute diseases	1,209	760	1,969	0.70	0.49	0.60	524	1145	1,669	0.80	1.45	1.16
8. all acute diseases	4,421	5,843	10,264	2.56	3.75	3.12	1,940	2,901	4,841	2.97	3.66	3.36
9. all chronic diseases	9,125	5,840	14,965	5.28	3.74	4.55	5,475	9,125	14,600	8.38	11.52	10.10
10. all diseases	13,546	11,683	25,229	7.84	7.49	7.67	7,415	12,026	19,441	11.35	15.18	13.46



TABLE 01.6. PERCENTAGE DISTRIBUTION OF DISEASES OR INJURIES ACCORDING TO TYPE OF TREATMENT RECEIVED FOR TWO SUB-SAMPLES

type of treatment	rural						urban					
	acute			chronic			acute			chronic		
	sub-sample 1	sub-sample 2	com-bined	sub-sample 1	sub-sample 2	com-bined	sub-sample 1	sub-sample 2	com-bined	sub-sample 1	sub-sample 2	com-bined
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1. allopath	35.98	42.80	39.42	46.77	46.67	46.72	50.72	33.85	42.54	60.53	95.56	79.52
2. homeopath	17.57	14.81	16.18	14.52	8.33	11.48	26.09	21.54	23.88	10.53	13.33	12.05
3. ayurved or unani	8.79	3.70	6.22	11.29	8.33	9.84	5.07	1.54	3.36	10.53	4.44	7.23
4. quack or no treatment	40.59	39.09	39.83	40.32	41.67	40.98	23.91	43.85	33.58	26.32	24.44	25.30
5. total	102.93 <sup>1</sup>	100.40	101.65	112.90	105.00	109.02	105.79	100.78	103.36	107.91	137.77	124.10
	(239) <sup>2</sup>	(243)	(482)	(62)	(60)	(122)	(138)	(130)	(268)	(38)	(45)	(83)

<sup>1</sup> Percentages will add up to more than 100, as some cases received more than one type of treatment.<sup>2</sup> Figures in parentheses are the numbers of cases reported during the reference period.



TABLE 01.7. MEDICAL EXPENDITURE (IN RUPEES) INCURRED PER ILLNESS DURING THE REFERENCE PERIOD ACCORDING TO NATURE OF ILLNESS AND TYPE OF TREATMENT FOR TWO SUB-SAMPLES

type of treatment	rural						urban					
	acute			chronic			acute			chronic		
	sub-sample I	sub-sample 2	com- bined	sub-sample I	sub-sample 2	com- bined	sub-sample I	sub-sample 2	com- bined	sub-sample I	sub-sample 2	com- bined
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1. allopath	8.02 (86) <sup>1</sup>	9.57 (104)	8.87 (190)	30.70 (29)	43.30 (28)	36.89 (57)	29.30 (70)	28.42 (44)	28.96 (114)	68.31 (23)	79.96 (43)	75.90 (66)
2. homeopath	5.57 (42)	3.58 (36)	4.65 (78)	20.97 (9)	6.16 (5)	15.79 (14)	17.83 (36)	7.82 (28)	13.45 (64)	61.88 (4)	34.38 (6)	45.38 (10)
3. ayurved or unani	4.78 (21)	6.41 (9)	5.30 (30)	47.07 (7)	13.85 (5)	33.23 (12)	87.14 (7)	55.50 (2)	80.11 (9)	41.25 (4)	28.35 (2)	37.95 (6)
4. quack or no treatment	1.52 (97)	1.50 (95)	1.51 (192)	1.25 (25)	4.33 (25)	2.79 (50)	4.30 (33)	9.00 (57)	7.28 (90)	8.91 (10)	6.01 (11)	7.39 (21)
5. total	4.90 (239) <sup>2</sup>	5.45 (243)	5.18 (482)	23.22 (62)	23.70 (60)	23.46 (122)	24.96 (138)	16.10 (130)	20.66 (268)	54.55 (38)	83.72 (45)	70.43 (83)

<sup>1</sup> Figures in parentheses are the numbers of cases on which the estimates are based.<sup>2</sup> Totals will not tally as some cases received more than one type of treatment.



# A PILOT HEALTH SURVEY IN WEST BENGAL—1955

## APPENDIX 2

### INDIAN STATISTICAL INSTITUTE

#### A PILOT HEALTH SURVEY IN WEST BENGAL: MARCH-MAY 1955: HOUSEHOLD SCHEDULE 1.1

##### Instructions to Investigators

- 03.1. (Block 2): Item 3. Household means of livelihood—As in industry—occupation code list (six digit code)
- 03.11. Item 4. The monthly expenditure per capita is to be worked out by first ascertaining the monthly household expenditure on consumer goods and dividing it by the number of members of the household.
- 03.12. Item 5. Religion—Hindu-0; Muslim-1; Sikh-2; Christian-3; Tribal-4; Others-5.
- 03.13. Item 6. Mother tongue—Bengali-0; Hindi-1; Urdu-2; Nepali-3; Tribal-4; Punjabi-5; Others-6.
- 03.14. Item 7. Purdah code—If women in the household do not observe purdah, enter code-1, and if they do, enter code-2.
- 03.15. Item 8. Informant's relation to head—head-0; spouse-1; son-2; daughter-3; father-4; mother-5; brother-6; sister-7; other relation-8; non-relation (household member)-9; others-10.
- 03.16. Item 9. Informant's ability—poor-0; average-1; good-2.
- 03.17. Item 10. Informant's willingness—hostile-0; unwilling-1; indifferent-2; helpful-3.
- 03.2. (Block 3): For each of the four visits enter date and signature.
- 03.3. (Block 4): Enter quantities consumed during the last 30 days in seers for items like rice, wheat, other cereals, ghee, oil, sugar and gur, milk, meat and fish (if the quantity is less than a seer enter-1). For consumption of eggs enter number and for fruits and vegetables enter values in rupees and annas. If consumption of the latter two items is from home production impute values at current local prices.
- 03.4. (Block 5): Item 1. Type of house—code-1—pucca house with brick walls, code-2—all other types of houses.
- 03.41. Item 2. Number of rooms—includes all living rooms and excludes those used for bath, cooking and store.
- 03.42. Item 3. Floor space—space under living rooms as well as those covered by verandahs if they are used for the same purpose as the living rooms. Give the figure in square feet.
- 03.43. Item 4. Ventilation—If the smoke has no good outlet and there is no possibility of free circulation of air, enter code-1; otherwise enter code-2.
- 03.44. Item 5. Water (drinking, washing)—The codes are same for drinking as well as washing water.  
code 1—tap water, code 2—tubewell water,  
code 3—well water, code 4—other types of water.
- 03.45. Item 6. Latrine code—  
code 1—sanitary privy, code 2—service privy,  
code 3—pit privy, code 4—others
- 03.46. Item 7. General sanitation—  
code 1—surroundings clean, drainage good and open space  
code 2—surroundings clean, either drainage is not good or lacks open space,  
code 3—surroundings unclean, covered with garbage and flies.



03.5. (Block 6) : Who are the members of the family?

- (a) All persons who have lived in the household and eaten from the household kitchen for at least 16 days during the month preceding the date of survey.
- (b) All children born within 14 days prior to the date of visit to members of the category (a).
- (c) All persons dead during the period, 14 days prior to date of visit, who if alive, would have been classed as (a).

03.51. Column 2—Relationship—(three-digit code)—head-0; spouse-1; son-2; daughter-3; father-4; mother-5; brother-6; sister-7; other relation-8; non-relation-9.

03.52. Column 3—Sex—male-1; female-2.

03.53. Column 4—Age—age last birthday.

03.54. Column 5—Marital status—never married-1; spouse living but divorced or separated-2; married-3; spouse dead-4.

03.55. Column 6—Nature of stay—For all persons present in this household throughout the whole year preceding the date of visit enter code-1. If the person has not stayed for the whole year ask—(i) whether present on the date of survey, (ii) whether stayed in the household for most of last fortnight and (iii) whether stayed in the household for most of last year and enter codes as follows :

code	present on date of visit	stayed for most of last fortnight	stayed for most of last year
2	yes	yes	yes
3	yes	yes	no
4	yes	no	yes
5	yes	no	no
6	no	yes	yes
7	no	yes	no
8	no	no	yes
9	no	no	no

For children below 1 year of age, enter code 1, if ever since birth, they were in this household.

03.56. Column 7—Educational status—(two digit code)—

*Left hand digit—education, general :*

Illiterate-1; literate but below primary-2; primary-3; middle-4; matric-5; intermediate-6; graduate in science-7; graduate in arts-8; post-graduate in science-9; post-graduate in arts-0.

*Right-hand digit—education, technical :*

no technical or professional qualification-1; technical or professional skill only, without degree or equivalent diploma but with or without certificate or diploma of lower order-2; holder of equivalent degree or diploma in teaching-3; engineering-4; agriculture-5; medicine, allopathic-6; other medicine-7; veterinary-8; law and commerce-9; other technology or profession-0.

03.57. Columns 14 and 15—Principal means of livelihood—As in industry-occupation code list (six digit code)

03.58. Column 18—Weight code—code 1—constantly losing weight, code 2—not constantly losing weight.

03.59. Column 19—Temperature code—code 1—constantly feeling feverish or rise of temperature.

03.5.10. Column 20—Health code—ask whether the person is constantly feeling fatigued and constantly lacking appetite and enter code as follows :

	<i>constant fatigue</i>	<i>constant lack of appetite</i>
code 1	yes	
code 2	yes	yes
code 3	no	no
code 4	no	yes
		no



## A PILOT HEALTH SURVEY IN WEST BENGAL—1955

- 03.6. Blocks, 1, 2 and 4 to 6 need be filled up only in the first visit. However, in subsequent visits some alterations might have to be made in block 6 for additions and exits of household members. Such entries may be made with an asterisk and footnotes given.
- 03.7. (Block 7): This has to be filled up only for such members of the household who were sick during the reference period. If a person fell sick more than once, for each sickness of the same person a separate line of entries is to be made. Cases of child birth, though medically treated should not be considered as cases of sickness.
- 03.71. Column 1—visit no.—for each sickness during the reference period of the 1st visit, enter 1 and for each sickness during the 2nd reference period enter 2 and so on.
- 03.72. Column 2—If any sickness was prevailing in the individual in this as well as in the previous reference period enter code 2. If it is a new case not prevailing during the previous visit enter code 1.
- 03.73. Column 3—Enter serial number of the affected person as entered in block 6. If he falls ill more than once during the reference period repeat his number for every sickness of his.
- 03.74. Column 4—Put down name of the disease as stated by the informant. If further particulars of the disease are given by the informant of his own accord, they may be entered in column 15 meant for 'remarks'.
- 03.75. Column 5—Some diseases are chronic i.e., they last for a long period of time and have no abrupt time of onset. Examples are T.B., heart diseases, diabetes, asthma, etc. Some diseases are acute and are of shorter duration with an abrupt time of onset and recovery if the result is not death. Examples are malaria, typhoid, cholera, dysentery, etc. It is also possible for certain diseases like malaria and dysentery to manifest as either acute or chronic. Another aspect of a disease is the degree of disability it causes on the stricken individual. If the affected person is not disabled from performing the usual assignment of work the disease is non-disabling. If it prevents him from doing his usual work, it is called disabling, the latter in extreme case may necessitate confinement to bed or hospital. In the case of old men, women and children not going to school it may be difficult to distinguish disability from non-disability unless the former is of a degree needing confinement to bed or hospital. For these people disability simply means taking of medicine or special diet. Based on these two aspects of the disease 6 codes for the nature of disease are given below:
- Chronic* : non-disabling—code 1; disabling but not confined to bed or hospital—code 2; confined to bed or hospital—code 3.
- Acute* : non-disabling—code 4; disabling but not confined to bed or hospital—code 5; confined to bed or hospital—code 6.
- 03.76. Column 6—Date of onset is the date on which disability starts—for all non-disabling diseases (acute or chronic) insert a dash in this column. For non-working persons date of onset is the date on which medical treatment or special diet starts.
- 03.77. Column 7—Date of recovery is the date on which disability ceases. If the person recovered on 15th March enter R—15th March and if he died on 15th March, enter D—15th March, the letter R or D preceding the date indicating the result of the disease. If the illness prevailed on the date of visit enter code 'P'.
- 03.78. Column 8—Period of sickness to be entered in terms of months, days.
- 03.79. Column 9—Code for type of attendance. If the case was attended by an allopath enter code 1 and if attended by homeopath or ayurved or unani or quack enter codes 2, 3, 4 or 5 respectively. If attended by none enter code 6.
- 03.7.10. Columns 10 to 12—All fees paid to physicians such as those for consultations, visits, operations, etc., are to be lumped up and entered in col. 10 and all expenses on medicines to be entered in col. 11 and all other expenses such as those for tonics, hospital rent, fees paid to



nurses and *dhais* etc., to be entered in col. 12. If the amount expended by a certain household is to cover more than one case of sickness it is necessary to allocate the respective proportion to the individual cases of sickness.

03.7.11. Column 13—Why medical care was not availed?

codes 1—no hospital or private physician available, 2—too expensive; 3—no faith in treatment; 4—sickness not serious; 5—other reasons.

03.7.12. Column 14—Reference period: for first visit the reference period is always 14 days. But for subsequent visits the actual number of days reckoned from the date of last visit to this one must be entered in the column.

03.8. (Block 8): The entries in this block pertain to only those women who have been delivered during one of the reference periods. As the survivalship of children born during any reference period has to be observed till the termination of the entire period of survey it is necessary to enter this item of information again in visits subsequent to the one in which the live birth was noted.

03.81. Columns 1 to 3: as in block 7.

03.82. Column 4—There are three types of termination. The codes are as follows: code 1—live birth, code 2—still birth and code 3—abortion.

03.83. Column 5—Date of delivery should be entered irrespective of the nature of termination.

03.84. Columns 6 to 8—If entry in col. 4 is code 1, then enter sex code of the child in col. 6, survival code (surviving—1, dead—2, left the household and not likely to return before the end of survey—3, left the household and is likely to return before the end of survey—4) in column 7, and age in (months, days) at present or at death or at departure in col. 8. For columns 7 and 8 entries are required in subsequent visits also.

03.85. Column 9—Place of delivery, code 1—delivered in this household, code 2—delivered in another household within district, code 3—delivered in hospital within district, code 4—delivered in another household outside district, code 5—delivered in hospital outside district.

03.86. Column 10—Attendance type: code 1—doctor, 2—midwife or nurse (qualified), 3—*dhai*, 4—hospital, 5—none, includes attendance by relatives and friends.

03.87. Column 11—Period of confinement—enter either period of hospitalisation or if home delivery enter period of bed-days. If on date of visit the woman is still lying on bed enter code 'bed' and in the next visit enquire again about the total bed-days and enter this item alongside of entries in columns 7 and 8.

03.88. Column 12—Period of convalescence—this means the period of disability following bed-days. If the woman is still convalescing enter code 'conv'. The method of entry is same as in col. 11.

03.89. Columns 13 to 16—Cost of medical care is split into four parts—column 13 gives expenses incurred towards physicians' fees for consultation, visit, operation etc. Column 14 gives fees paid to midwife or *dhai*. Column 15 gives expenditure on medicine and column 16 gives cost of hospitalisation.

03.8.10. Column 17—Reference period—enter as in block 7.

03.8.11. Column 18—Survival of mother—code 1—mother alive, code 2—mother dead.

03.8.12. In block 8 if any woman has given birth to twins two consecutive lines must be entered.

03.9. (Block 9): Every woman married, widowed, separated or divorced must have an entry in this block. Her serial number as in block 6 must be entered in column 1.

03.91. Column 2—Age at present should be copied from block 6.

03.92. Column 3—Age at marriage is the age at first marriage for women married more than once.

03.93. Column 4 to 9—Enter codes for educational status, economic status and means of livelihood for the woman and her last husband.



# A PILOT HEALTH SURVEY IN WEST BENGAL—1955

03.94. Columns 11 to 33—These give the age of the mother and the result of termination for successive orders of terminations. If the last termination occurred within last one year the age of the mother and the result of termination should not be given in the columns appropriate to its order but columns 31 to 33 must be entered. Column 31 gives the order of this last termination, column 32 gives the calendar month in which the termination took place and column 33 are, code 1—live birth, child surviving on date of visit, code 2—still birth, code 3—abortion, and if the result was a live birth but the child is dead then enter 'D-N' where N stands for the completed months of life (say, if the child died after 3 months of life enter 'D-3'). If the last termination took place at least one year ago, then entries in columns 31 to 33 should not be made. Such a termination should be entered as the previous ones. Against age give the age of mother at the time of termination and against result enter codes as follows:

code 1—live birth, and child survived first year of life,  
code 2—live birth, child died within first month of life,  
code 3—live birth, child died after one month but before one year of its life,  
code 4—still birth, code 5—abortion.

Example 1: (1) Woman's present age—35, (2) age at marriage—18, (3) total terminations—2, (4) 1st termination occurred at age 20, child died at 6 months of life, (5) 2nd termination occurred at age 22 and resulted in still birth.

Since the woman's age at present is 35 and her last termination occurred at 22, i.e., 13 years ago no entries are needed in columns 31–33.

The entries are:

col. 2	col. 3	col. 10	col. 11	col. 12	col. 13	col. 14	col. 31	col. 32	col. 33
35	18	2	20	3	22	4	—	—	—

Example 2: (1) woman's age at present 28, (2) age at marriage 20, (3) total terminations 4, (4) 1st termination at age 22, live birth child survived first year of life, (5) 2nd termination at age 24, child died before 1 month of life. (6) 3rd termination at age 26, still birth, (7) 4th termination at age 28, child died after one month but before 1 year of life. Since the last termination occurred at age 28 and the woman's present age being 28 this termination must have occurred within last year. Ask for the calendar month of this termination and the age at death in completed months. Suppose the answer is 'May, 1954' and age at death is 6 months. Then the entries are:—

col. 2	col. 3	col. 10	col. 11	col. 12	col. 13	col. 14	col. 15	col. 16	col. 17	col. 18	col. 31	col. 32	col. 33
28	20	4	22	1	24	2	26	4	—	—	4	May	D-6

Column 34—Information on ante-natal care is to be collected only in respect of terminations occurring during the last 1 year.

Two-digit code—left hand digit indicating the type of attendance—  
no attendance—0; hospital—1; welfare centre—2; qualified practitioner—3; qualified midwife—4.

Right-hand digit—number of such attendance.  
If there is more than one type of attendance, only the predominant type need be recorded.







[illegible]



(8) current terminations—reference period 2 weeks for 1st visit and period since last visit for all subsequent visits.

(9) history of terminations



# ON RECALL LAPSE IN INFANT DEATH REPORTING<sup>1</sup>

By RANJAN KUMAR SOM

and

NITAI CHANDRA DAS

Indian Statistical Institute, Calcutta

**SUMMARY.** Recall lapse as a distorting factor in infant death (and sex ratio at birth) reporting in historical studies, based on the interview method at a current moment, was introduced in the Indian demographic situation by Mahalanobis and Das Gupta (1954) and elaborated by Das Gupta, Som, Majumdar, and Mitra (1955) in *Couple Fertility*; the approach in these two studies, based on the National Sample Survey data, was through the marriage cohorts where time entered as a distinct component. The trend of the observed proportions over time was seen to get substantially distorted from that of the true proportions, obtained from external evidences, establishing thus the existence of a definite, progressively decreasing recall lapse in infant death reporting.

Poti, Raman, Biswas, and Chakravarty (1959), analyzing the data on infant death of the West Bengal Household Comparative Study and Health and Employment Study 1955, by order of birth concluded that recall lapse in infant death reporting was not statistically significant. The present paper seeks to establish the theoretical validity of the study of the recall lapse in infant death reporting by the marriage cohort differential vis-a-vis that by the order of birth differential and analyzing the same data as that utilized by Poti etc., by marriage cohorts, confirms the findings of the previous two studies.

1. Recall lapse as a constituent, distorting factor in infant death (and sex ratio at birth) reporting in historical studies based on the interview method at a current moment was first introduced by Mahalanobis and Das Gupta (1954) for the Indian demographic situation, obtained from the National Sample Survey (NSS). In *Couple Fertility* by Das Gupta, Som, Majumdar, and Mitra (1955), based also on the same NSS data, this problem was studied in greater details and tentative exponential curves fitted, with good agreement, for the percentages under-reporting of infant deaths by marriage cohorts.

2. Let  $m_t$  mothers married at calendar year  $t$  have  $b_{ti}$   $i$ -th order of births, of which  $d_{ti}$  die in their first year of life; then  $p_{ti}$ , the infant death proportion (IDP) of the  $i$ -th order of birth to such mothers is

$$p_{ti} = d_{ti}/b_{ti}.$$

In a survey, where the past fertility history is collected through interview at a current moment, defective number of births and infant deaths  $b'_{ti}$  and  $d'_{ti}$  respectively are reported, giving the corresponding IDP

$$p'_{ti} = d'_{ti}/b'_{ti}.$$

The problem is to find out if recall lapse operates in infant death reporting so that the trend of the observed proportions ( $p'$ ) over time might get substantially distorted from that of the true proportions ( $p$ ).

3. In the study by Mahalanobis and Das Gupta and also in *Couple Fertility*, this topic was studied by the marriage cohort differential. The observed IDP for the  $m_t$  mothers over all birth orders is

$$\sum_i d'_{ti} / \sum_i b'_{ti} = \sum_i b'_{ti} p'_{ti} / \sum_i b'_{ti} = \bar{p}'_t, \text{ say.}$$

A comparison of  $\bar{p}'_t$  with  $\bar{p}_\tau$  ( $\tau$  being more recent than  $t$ ) was then made to show that significant recall lapse existed in infant death reporting, since the observed proportions gave the relation  $\bar{p}'_\tau > \bar{p}'_t$  while from external evidence it could be stated that  $\bar{p}_\tau < \bar{p}_t$ . (In the above two studies, the IDP was analyzed for marriage cohort groups and not for any

<sup>1</sup>A rejoinder to "A pilot health survey in West Bengal", by Poti, Raman, Biswas and Chakravarti (1959), *Sankhyā*, 21, 141-204.



particular marriage cohorts; we can, however, assume without any loss in generality that  $t$  and  $\tau$  are the central points of two marriage cohort groups).

4. The problem of recall lapse in infant death reporting has recently been referred to by Poti, Raman, Biswas, and Chakravarti (1959) in "A pilot health survey in West Bengal, 1955", based on the data of the West Bengal Household Comparative Study and Health and Employment Study, 1955<sup>1</sup> published in this issue on p. 141-204. Here the study of the recall lapse in infant death reporting was limited to the different birth orders from mothers having five or more terminations. The observed IDP for the  $i$ -th birth order over all mothers (and marriage cohorts) is

$$\sum_t d'_{ti} / \sum_t b'_{ti} = \sum_t b'_{ti} p'_{ti} / \sum_t b'_{ti} = \bar{p}'_{\cdot i}, \text{ say.}$$

Thus, a comparison between the IDP's  $\bar{p}'_{\cdot i}$  and  $\bar{p}'_{\cdot i}$  for two birth orders  $i$  and  $i$  would not reflect the effect of recall lapse over time (apart from the small interval between successive births which is of the order of only 2-3 years). The observed IDPs by orders of birth for the West Bengal Health Survey can not then, as the authors made them out on the basis of their Table 4.1, contradict the existence of a substantial recall lapse over time between marriage cohorts, observed in the previous two studies on a national scale.

5. From the same data as that of Table 4.1 of "A pilot health survey in West Bengal", the IDPs for the different marriage cohort groups calculated by us are presented in Table 1. This table, which has a counterpart in Table 8 of Mahalanobis and Das Gupta (1954) and Table 8.1 of *Couple Fertility* (Das Gupta, Som, Majumdar, and Mitra, 1955), on the other hand shows clearly the existence of similar recall lapse over time in infant death reporting except for small kinks (occasioned perhaps by the small sample sizes), and confirms the findings of Mahalanobis and Das Gupta (1954) as also of Das Gupta, Som, Majumdar, and Mitra (1955).

TABLE 1. INFANT DEATH PROPORTION (PER 1000 LIVE BIRTHS) FOR EVER-MARRIED WOMEN HAVING FIVE OR MORE TERMINATIONS BY MARRIAGE COHORT GROUPS: WEST BENGAL HOUSEHOLD COMPARATIVE STUDY AND HEALTH & EMPLOYMENT STUDY, 1955

marriage cohort	rural	urban
(1)	(2)	(3)
1. before 1910	147	99
2. 1910-19	155	158
3. 1920-29	200	180
4. 1930-39	183	115
5. 1940-45	188	221
6. all marriage cohorts	169	140
(no. of mothers)	(522)	(165)

<sup>1</sup> In this survey the approach was through the living married women which would, on the basis of the findings of *Couple Fertility*, miss about 12 per cent of the broken couples with wife dead. The effect of the differing approach adopted on the recall lapse in infant death reporting is, however, not being discussed here.



## ON RECALL LAPSE IN INFANT DEATH REPORTING

6. In "A pilot health survey in West Bengal", the recall lapse in infant death reporting was also studied for mothers aged 43 years or over, the births to these mothers being divided into two groups—(i) births occurring within 15 years preceding the date of survey; and (ii) births occurring before 15 years preceding the date of survey. The IDP was seen to be greater for birth group (ii) than that for (i), from which also the authors concluded that recall lapse was not statistically significant.

7. The same data on which the above finding was based (Table 4.2 of "A pilot health survey in West Bengal") were analyzed in this note by marriage cohorts. For births occurring within 15 years preceding the date of survey, the sample sizes were inadequate for the individual marriage cohort groups. For births occurring 15 years or earlier preceding the date of survey, the IDPs are presented in Table 2 by marriage cohort groups. From this table, it will be seen that the analysis by marriage cohort shows up the existence of recall lapse for these mothers also. The division of the births for mothers aged 43 years or over into two groups by birth period and analysis over all marriage cohorts within each birth period group is not expected to show up the effect of recall lapse in infant death reporting as for the first group, i.e., for births occurring within 15 years preceding the date of survey, about 9 per cent in rural areas and 12 per cent in the urban relate to the first four births : the corresponding proportion for births occurring 15 years or earlier preceding the date of survey is 63 per cent in rural areas and 71 per cent in the urban. Thus the second group, being loaded heavily in favour of the earlier orders of births, just presents again the finding that the observed IDPs for the earlier orders of birth in this particular study were higher than those for the later orders; this was, in fact, a general feature which ran through all marriage cohorts for all mothers.

TABLE 2. INFANT DEATH PROPORTION (PER 1000 LIVE BIRTHS) FOR BIRTHS OCCURRING 15 YEARS OR EARLIER PRECEDING THE DATE OF SURVEY TO EVER-MARRIED WOMEN AGED 43 YEARS OR OVER, BY MARRIAGE STUDY COHORT GROUPS: WEST BENGAL HOUSEHOLD COMPARATIVE & HEALTH AND EMPLOYMENT STUDY, 1955

marriage cohort	rural	urban
(1)	(2)	(3)
1. before 1910	140	106
2. 1910-19	172	175
3. 1920-29	180	145
4. all marriage cohorts <sup>1</sup>	160	137
(no. of mothers)	(359)	(153)

<sup>1</sup> Including the marriage period 1930-39 with a very small sample size.



8. The fallacy which crept in the arguments by the authors of "A pilot health survey of West Bengal" was inherent in the assumption that recall lapse in infant death reporting was order of birth-specific, and not mother-(or marriage cohort-) specific. In the previous two studies, the same group of mothers did not appear more than once in the differentials studied whereas in this particular study they do and vitiate the homogeneity of data. For either substantiation or invalidation of the phenomenon of recall lapse observed in the former studies, the same line of analysis should have been followed here also : even then, a regional survey cannot always show up the same features as may be observed in a study on a national scale, which is expected to smooth out regional peculiarities if they exist.

## REFERENCES

- DAS GUPTA, AJIT; SOM, RANJAN KUMAR; MAJUMDAR, MURARI; AND MITRA, SAMARENDRA NATH (1955): Couple Fertility. *National Sample Survey Number 7*, Government of India, Ministry of Finance, Department of Economic Affairs & *Sankhyā*, **16**, 230-434.
- MAHALANOBIS, P. C. AND DAS GUPTA, AJIT (1954): The use of sample surveys in demographic studies in India. *UN World Population Conference, Rome, 1954, E/Conf. 13/194*.
- POTI, S. J., RAMAN, M. V., BISWAS, S., AND CHAKRAVARTI, B. (1959): A pilot health survey in West Bengal 1955, *Sankhyā*, **21**, 141-204.

*Paper received : May, 1958.*



# SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS

1. PLACE OF PUBLICATION : 204/1 Barrackpore Trunk Road, Calcutta-35
2. PERIODICITY OF ITS PUBLICATION : Quarterly
3. PRINTER'S NAME : Kalipada Mukherjee  
NATIONALITY : Indian  
ADDRESS : 204/1 Barrackpore Trunk Road, Calcutta-35
4. PUBLISHER'S NAME : Prafulla Chandra Mahalanobis  
NATIONALITY : India  
ADDRESS : 204/1 Barrackpore Trunk Road, Calcutta-35
5. EDITOR'S NAME : P. C. Mahalanobis  
NATIONALITY : Indian  
ADDRESS : 204 Barrackpore Trunk Road, Calcutta-35
6. NAMES AND ADDRESSES OF INDIVIDUALS  
WHO OWN THE NEWSPAPER AND PART-  
NERS OR SHAREHOLDERS HOLDING MORE  
THAN ONE PER CENT OF THE TOTAL  
CAPITAL : Published by Statistical Publishing Society  
(non-profit distributing agency)

I, Prafulla Chandra Mahalanobis, hereby declare that the particulars given above are true to the best of my knowledge and belief.

*Date. 26 March, 1959*

*Sd/ Prafulla Chandra Mahalanobis*







# SANKHYĀ

## THE INDIAN JOURNAL OF STATISTICS

*Edited by : P. C. MAHALANOBIS*

---

VOL. 21, PARTS 3 & 4

AUGUST

1959

---

### THE ANALYSIS OF HETEROGENEITY. I.

*By J. B. S. HALDANE*

*Indian Statistical Institute, Calcutta*

**SUMMARY.** Estimators of the mean and variance of a frequency are given when this frequency varies through a series of samples.

#### INTRODUCTION

The following situation frequently arises in biological research, and doubtless in other branches of science. A number of experiments or observations are made under as nearly similar conditions as possible. Each of them leads to the production of a sample, whose members may be classed into two types which we may call successes and failures, though they may be females and males, fertile and sterile matings, survivals and deaths, and so on. If we have  $n$  samples, and the  $i$ -th consists of  $s_i$  members of which  $a_i$  are successes and  $b_i$  are failures, we can draw up a  $(2 \times n)$  fold table and apply the  $\chi^2$  or some other test of homogeneity. If the test is judged compatible with homogeneity, we can adopt the simple hypothesis that the probability of success was the same in each sample, and estimate it as  $p = \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n s_i}$ . If however the test is judged significant of heterogeneity, we must conclude that  $p$  has varied from one experiment to another, its value in the  $i$ -th experiment being  $p_i$ . We can then proceed to estimate the mean of  $p_i$ . Although the estimate given above is unbiased, it will be shown that it is not efficient unless the sample number  $s_i$  is constant. We can also in general estimate the variance and higher moments of  $p_i$ . While  $\chi^2$  is a test for heterogeneity it is not a measure of it, but we shall see that the estimate of the variance of  $p$  is related to  $\chi^2$ . In this paper I shall only deal with the estimation of the mean and variance.

In my experience this problem has arisen in two rather different contexts. On the one hand we may have to analyse a series of litters of mice or other small



animals produced from parents as similar as possible, under standardised conditions. The mean value of  $s$  is of the order of 6, and for most values of  $s$  there is a fair number of samples. If heterogeneity has been detected, we can estimate the mean value of  $p$  for various sample numbers, and see whether they regress significantly on  $s$ . If they do not, we can weight each with the appropriate amount of information, and combine them. If our data are numerous enough, we can do the same for other moments.

On the other hand we may have to deal with a series of insect or plant families, in which the mean value of  $s$  is about 100, and two samples with the same value of  $s$  are unusual.

As Robertson (1951) pointed out, this problem is the inverse of the problem studied by Lexis (1877). Lexis considered the effect on the variance of  $a_i$  of a known variance of  $p_i$ .

In what follows I use  $\kappa_r$  to mean the  $r$ -th cumulant of the true distribution of  $p$ .  $k_r$  means an unbiased estimate of  $\kappa_r$  and  $\kappa(r^s)$  the expectation of the  $s$ -th cumulant of the distribution of  $k_r$ , while  $k(r^s)$  is an unbiased estimate of  $\kappa(r^s)$ . We have to consider expectations at two levels. I denote expectations for a given value of  $p_i$ , and thus within a single sample, with an asterisk. Thus  $\mathcal{E}^*(a_i) = p_i s_i$ ,  $\mathcal{E}^*(a_i^2) = p_i^2 s_i(s_i-1) + p_i s_i$ ,  $\mathcal{E}^*(a_i b_i) = p_i(1-p_i) s_i(s_i-1)$  and so on. I denote expectations within the whole group of  $n$  samples without an asterisk. Thus  $\mathcal{E}(p) = \mathcal{E}(p_i) = \kappa_1$ . I use  $\Sigma a$  or  $\Sigma a_i$  to mean  $\sum_{i=1}^n a_i$ , and so on. If  $s$  is constant  $\mathcal{E}(\Sigma a) = \kappa_1 ns$ . If  $s$  is variable I assume that  $s_i$  and  $p_i$  are uncorrelated, though this should be verified where possible. In any case I assume that  $p_i$  and  $p_j$  are uncorrelated, that is to say  $\mathcal{E}(p_i p_j) = \kappa_1^2$ , if  $i \neq j$ . Also

$$\mathcal{E}(p_i) = \kappa_1, \mathcal{E}(p_i^2) = \kappa_1^2 + \kappa_2.$$

#### SAMPLES OF CONSTANT SIZE

If every sample consists of  $s$  members, then since  $\mathcal{E}^*(a_i) = p_i s$ , so  $\mathcal{E}(\Sigma a) = \kappa_1 ns$ , whence

$$k_1 = (ns)^{-1} \Sigma a. \quad \dots (1)$$

This estimate is clearly unbiased and efficient.

$$(\Sigma a)^2 = \Sigma a^2 + 2 \sum_i \sum_{j \neq i} a_i a_j.$$

$$\begin{aligned} \text{So} \quad \mathcal{E}[(\Sigma a)^2] &= ns(s-1) \mathcal{E}(p^2) + ns \mathcal{E}(p) + n(n-1)s^2 \mathcal{E}(p_i p_j) \\ &= ns(s-1) (\kappa_1^2 + \kappa_2) + ns \kappa_1 + n(n-1)s^2 \kappa_1^2 \\ &= ns(ns-1)\kappa_1^2 + ns \kappa_1 + ns(s-1)\kappa_2. \end{aligned}$$

$$\text{But} \quad [\mathcal{E}(\Sigma a)]^2 = n^2 s^2 \kappa_1^2,$$

$$\text{so} \quad \text{var}(\Sigma a) = ns(\kappa_1 - \kappa_1^2) + ns(s-1)\kappa_2,$$

$$\text{whence} \quad \kappa(1^2) = \text{var}(k_1) = \frac{\kappa_1(1-\kappa_1) + (s-1)\kappa_2}{ns}. \quad \dots (2)$$



# THE ANALYSIS OF HETEROGENEITY. I.

The first term is the component due to the small sample size, while the second,  $n^{-1}(1-s^{-1})\kappa_2$ , is due to the variance of  $p$ . The second term may greatly exceed the first.

$$\sum a_i \sum b_i = \sum_i a_i b_i + \sum_i \sum_{j \neq i} a_i b_j.$$

$$\begin{aligned} \text{So } \mathcal{E}[\sum a_i \sum b_i] &= ns(s-1) (\kappa_1 - \kappa_1^2 - \kappa_2) + n(n-1)s^2(\kappa_1 - \kappa_1^2) \\ &= ns(ns-1)\kappa_1(1-\kappa_1) - ns(s-1)\kappa_2. \end{aligned}$$

$$\text{Also } \mathcal{E}[\sum a_i b_i] = ns(s-1) (\kappa_1 - \kappa_1^2 - \kappa_2).$$

$$\text{Hence } \mathcal{E}[(s-1)\sum a \sum b - (ns-1) \sum ab] = n(n-1)s^2(s-1)\kappa_2.$$

$$\text{So } k_2 = \frac{(s-1) \sum a \sum b - (ns-1) \sum ab}{n(n-1)s^2(s-1)} \quad \dots (3)$$

is an unbiased estimate of  $\kappa_2$ . Also

$$\mathcal{E} \left[ \frac{\sum a \sum b - \sum ab}{n(n-1)s^2} \right] = \kappa_1(1-\kappa_1).$$

So we can put (2) in terms of observed quantities.

$$k(1^2) = \frac{\sum a \sum b - n \sum ab}{n^2(n-1)s^2} = \frac{\text{Cov}(a, b)}{(n-1)s^2}. \quad \dots (4)$$

Robertson (1951) gave an expression for the variance of  $p$  which in my symbolism becomes :

$$k_2 = [n^2 s^2 (s-1)]^{-1} [(s-1)\sum a \sum b - ns \sum ab].$$

Unless  $n$  is small, this is very near to my expression (3), the difference being the value (4) of  $k(1^2)$ . However when  $n$  is small the difference is not negligible. For example if  $s = 100$ ,  $n = 5$ , and the values of  $a$  are 4, 8, 10, 14, 17, then  $\sum a = 53$ ,  $\sum b = 447$ ,  $\sum ab = 4635$ .  $k_1 = 0.106$ , and expression (3) gives .0016436 for the variance or .04054 for the standard deviation of  $p$ , while Robertson's expression gives .00112763 and .03358. If my own value is judged to be more accurate, it should be used.

$\chi_{n-1}^2$ , used as a test of homogeneity, may be written

$$\chi_{n-1}^2 = \frac{ns(\sum a \sum b - n \sum ab)}{\sum a \sum b}.$$

When  $\kappa_2 = 0$ , that is to say  $p$  is constant, its exact expectation is

$$\mathcal{E}(\chi_{n-1}^2) = (ns-1)^{-1} n(n-1)s.$$

$$\text{So } \chi_{n-1}^2 - (ns-1)^{-1} n(n-1)s = [(ns-1) \sum a \sum b]^{-1} n^3(n-1)s^3(s-1)k_2,$$

$$k_2 = \frac{\sum a \sum b [ (ns-1)\chi_{n-1}^2 - n(n-1)s ]}{n^3(n-1)s^3(s-1)}. \quad \dots (5)$$

or



Since  $\chi_{n-1}^2$  can be zero with a finite probability, but cannot be negative, it follows that  $k_2$  can be negative, its minimum value, if  $p$  is constant, being  $-p(1-p)(s-1)^{-1}$ . The null sampling distribution of  $k_2$  is given by that of  $\chi_{n-1}^2$ , and the significance of a positive value is that of the corresponding value of  $\chi^2$ . If  $k_2$  is negative we must suppose that  $\kappa_2$  is zero or too small to estimate, while drawing the appropriate conclusions from the small value of  $\chi_{n-1}^2$ .

We now see that  $\chi^2$ , as a test of homogeneity, has a triple function. Firstly its excess over its null value furnishes a test of whether the variance of  $p$  exceeds zero. Secondly it allows us, by means of (5), to estimate the variance of  $p$ . And thirdly it measures the uncertainty of the estimate of the mean of  $p$ . For (4) may be written as

$$\text{var}(k_1) = [n^3(n-1)s^3]^{-1} \chi_{n-1}^2 \Sigma a \Sigma b. \quad \dots (6)$$

Workers are rightly suspicious of a mean based on a heterogeneous set of samples. (4) or (6) tells them just how suspicious they should be. I may add that if  $\chi_{n-1}^2$  is calculated by the method of Haldane (1955) which, it is claimed, saves a good deal of computation in some cases, (5) and (6) are more useful than (3) and (4).

(5) is analogous to the well-known relation between  $r$  and  $\chi^2$  for a  $(2 \times 2)$ -fold table. I hope to give estimators of  $\kappa_3$ ,  $\kappa_4$ , and of the sampling variance of  $k_2$ , in a later paper. The latter is not however of immediate importance, since we have an expression for the significance of a given value of  $k_2$ .

#### SAMPLES OF VARIABLE SIZE

If we have a large number of samples for each of a few small values of  $s$ , as with human families, we can treat each set for a given value of  $s$  separately, and combine the estimates of  $p$ , the amount of information for each value of  $s$  being given by (2). When however, the values of  $s$  are all or mostly different, we proceed as follows.

If  $w_i$  be any weighting factor, then,

$$\mathcal{E}(\Sigma w_i a_i) = \kappa_1 \Sigma w_i s_i.$$

So provided  $\Sigma w_i s_i = 1$ ,  $\Sigma w_i a_i$  is an unbiased estimate of  $\kappa_1$ . Clearly  $w$  must be a one-valued function of  $s$ . Clearly also when  $\kappa_2 = 0$ , that is to say  $p$  is constant,  $w_i$  should be constant, and therefore equal to  $(\Sigma s)^{-1}$ . When however  $\kappa_2$  is not zero,  $w$  should be an increasing function of  $s$ . One can derive the expression  $w_i \propto [s_i - 1 + \kappa_2^{-1} \kappa_1 (1 - \kappa_1)]^{-1}$  which follows, directly from (2) by a somewhat intuitive argument. But it can be derived more rigorously as follows.

The most efficient form of  $w_i$  is that which minimizes the variance of  $k_1 = \Sigma_i w_i a_i$ . Now

$$k_1^2 = \Sigma_i w_i^2 a_i^2 + 2 \Sigma_i \Sigma_{j \neq i} w_i a_i w_j a_j.$$



# THE ANALYSIS OF HETEROGENEITY. I.

Since  $\mathcal{E}[a_i^2] = (\kappa_1^2 + \kappa_2)s_i(s_i - 1) + \kappa_1 s_i$   
 and  $\mathcal{E}[a_i a_j] = \kappa_1^2 s_i s_j$ ,  
 it follows that  $\mathcal{E}[k_1^2] = \sum_i w_i^2 s_i [(s_i - 1)(\kappa_1^2 + \kappa_2) + \kappa_1] + 2 \sum_i \sum_{j \neq i} w_i w_j s_i s_j \kappa_1^2$   

$$= \kappa_1^2 (\sum w_i s_i)^2 + \kappa_1 (1 - \kappa_1) \sum w_i^2 s_i + \kappa_2 \sum w_i^2 s_i (s_i - 1).$$

But  $\sum w_i s_i = 1$ , so on subtracting  $\kappa_1^2$  we find

$$\text{var}(k_1) = \sum_i [\{\kappa_2 + (\kappa_1 - \kappa_1^2 - \kappa_2)s_i^{-1}\} w_i^2 s_i^2].$$

Since  $\sum_i w_i s_i = 1$ , this is minimal when  $w_i s_i \propto [\kappa_2 + (\kappa_1 - \kappa_1^2 - \kappa_2)s_i^{-1}]^{-1}$

or

$$w_i = [\kappa_2 s_i + \kappa_1 - \kappa_1^2 - \kappa_2]^{-1} [\{\kappa_2 + (\kappa_1 - \kappa_1^2 - \kappa_2)s_i^{-1}\}]^{-1}.$$

So if

$$c = \kappa_1(1 - \kappa_1)\kappa_2^{-1} - 1, \quad w_i \propto (s_i + c)^{-1},$$

and

$$k_1 = \frac{\sum \left( \frac{a}{s+c} \right)}{\sum \left( \frac{s}{s+c} \right)} \quad \dots \quad (7)$$

$$\text{var}(k_1) = \kappa(1^2) = \frac{\kappa_2}{\sum \left( \frac{s}{s+c} \right)}. \quad \dots \quad (8)$$

If  $\kappa_2$  is small, that is to say  $p$  is nearly constant,  $c$  is very large, and  $k_1$  approximates to  $(\sum s)^{-1} \sum a$ , as is obvious. If  $p$  can only assume the values of zero and unity,  $c = 0$ , and  $k_1 = n^{-1} \sum a s^{-1}$ . Otherwise the values of  $c$  are intermediate. Thus if all values of  $p$  from 0 to 1 are equally frequent,  $c = 2$ , and if all values from 0 to  $\frac{1}{2}$  are equally frequent,  $c = 20$ , and so on. Usually therefore it will be necessary to estimate  $\kappa_2$ .

Prof. C. R. Rao has pointed out to me that the estimate (7) is not quite unbiased, because in general estimates of  $\kappa_1$  and  $\kappa_2$  are correlated. However the bias will seldom be large. The most efficient estimator of  $\kappa_2$  can only be given when higher moments are known, so that formally the problem is very complicated. But an infinite number of unbiased estimators of  $\kappa_2$  can be derived, according to the weights given to different samples. The weight to be attached to any sample will always increase with  $s$ , but the weight as a function of sample size will be somewhere between that appropriate when  $s$  is large and  $\kappa_2$  small, and that appropriate when  $s$  is small and  $\kappa_2$  large. We can write down any number of expectations, including the following :

$$\begin{aligned} \mathcal{E}[\sum s^{-1} ab] &= (\kappa_1 - \kappa_1^2 - \kappa_2)(\sum s - n) \\ \mathcal{E}[\sum (s-1)^{-1} ab] &= (\kappa_1 - \kappa_1^2 - \kappa_2) \sum s \\ \mathcal{E}[\sum s^{-1} (s-1)^{-1} ab] &= (\kappa_1 - \kappa_1^2 - \kappa_2) n \\ \mathcal{E}[\sum a \sum b] &= \kappa_1(1 - \kappa_1)(\sum s - 1) \sum s - \kappa_2(\sum s^2 - \sum s) \\ \mathcal{E}[\sum s^{-1} a \sum s^{-1} b] &= \kappa_1(1 - \kappa_1)(n^2 - \sum s^{-1}) - \kappa_2(n - \sum s^{-1}). \end{aligned}$$



From these we can at once derive a number of unbiased estimates of  $\kappa_2$ , of which the most likely to be useful are :

$$\begin{aligned}
 k_{2\alpha} &= \frac{(\sum s - n)\sum a \sum b - (\sum s - 1)\sum s \sum (s^{-1}ab)}{(\sum s - n)[(\sum s)^2 - \sum s^2]} \\
 &= \frac{\sum a \sum b [(\sum s - 1)\chi_{n-1}^2 - (n-1)\sum s]}{(\sum s - n)[(\sum s)^2 - \sum s^2]} \quad \dots \quad (9)
 \end{aligned}$$

$$k_{2\beta} = \frac{\sum a \sum b - (\sum s - 1)\sum (s-1)^{-1}ab}{(\sum s)^2 - \sum s^2} \quad \dots \quad (10)$$

$$k_{2\gamma} = \frac{n\sum s^{-1}a \sum s^{-1}b - (n^2 - \sum s^{-1})\sum s^{-1}(s-1)^{-1}ab}{n^2(n-1)} \quad \dots \quad (11)$$

Of these estimates  $k_{2\alpha}$  and  $k_{2\beta}$  should differ very little, and  $k_{2\beta}$  is the easiest to compute unless  $\chi_{n-1}^2$  has already been computed, which however will usually be the case. It will be seen that  $k_{2\alpha}$  and  $k_{2\beta}$  have about the same weighting as  $\chi^2$ , while  $k_{2\gamma}$  assigns approximately equal weight to each sample.

From the example which follows it will be seen that these estimates may be very close to one another. Indeed  $k_{2\alpha}$  and  $k_{2\beta}$  agree to four significant figures. They thus furnish a fairly precise estimate of  $\kappa_2$ , which, in turn, allows an accurate estimate of  $\kappa_1$ , and of the sampling variance of this estimate of  $\kappa_1$ .

TABLE 1. RECOMBINANTS IN 13 CULTURES OF  
*DROSOPHILA SUBOBSCURA*

$s$	$a$	$b$	$p' = s^{-1}a$
224	69	155	.30804
206	59	147	.28641
255	70	185	.27451
267	70	197	.26217
247	61	186	.24696
238	57	181	.23950
166	36	130	.21687
199	42	157	.21106
210	39	171	.18571
284	50	234	.17608
190	33	157	.17368
187	32	155	.17113
243	40	203	.16461
2916	658	2258	2.91673



# THE ANALYSIS OF HETEROGENEITY. I.

## A NUMERICAL EXAMPLE

In Table 1 the successive values of  $s$  are the numbers of imagines of *Drosophila subobscura* in 13 bottles each derived from a single pair mating. In each bottle the father was homozygous for a pair of autosomal recessive genes belonging to the same linkage group and therefore located on the same chromosome. The mother was heterozygous at these two loci. The values of  $a$  are the numbers of flies in which these loci had undergone recombination, commonly described as "cross-overs." I have to thank Mrs Trent, of the Department of Biometry, University College, London, for these figures.  $b = s - a$ , and  $p'_i = a_i s_i^{-1}$ . That is to say the value of  $p'_i$  is the estimate of recombination frequency from the  $i$ -th culture. The cultures are arranged in descending order of  $p'_i$ .  $\chi^2_{12} = 37.059$ ,  $P(\chi^2) = .00022$ , so there is very strong evidence of heterogeneity.

Now if  $p$  were constant, its estimate would be  $658/2916 = .2257 \pm .0073$ . In fact all estimates known to me have been made in this way.

The mean value of  $p'_i$  is .2244, its median .2169.  $\text{var}(p'_i) = .002380$ , so  $\sigma_{p'} = .0487$ . This is much too high as an estimate of the variance of  $p$ . The formulae (9) and (11) give  $k_{2\alpha} = k_{2\beta} = .001636$ ,  $k_{2\gamma} = .001682$ . This is a very satisfactory agreement and we may estimate  $\kappa_2$  as .00166, giving  $\sigma_p = .0407$  which is considerably below the crude estimate of .0487.

Adopting the provisional value of .225 for  $\kappa_1$ , we find  $c = 105.03$ . Putting  $c = 105$  in (7),  $k_1 = .2273$ . If we repeat the process we find  $c = 106.3$ , which does not alter the value of  $k_1$ . From (8) we find  $\kappa_2(1^2) = .0001908$ . So

$$k_1 = .2273 \pm .0138.$$

Thus the estimate of the mean of  $p$  is only changed from its "classical" value by 12% of its standard sampling error. The change could be much greater if the values of  $s$  had a larger coefficient of variation. On the other hand its standard sampling error is nearly doubled. And we have at least an estimate of the variance of  $p$ .

## DISCUSSION

This paper is a preliminary attempt to develop a field of statistics opened up by Robertson (1951). If the sample size  $s$  is constant it is merely a matter of algebraical accuracy to obtain unbiased and efficient estimates of all the moments or cumulants of the distribution of  $p$ , upto and including the  $s$ -th. On the other hand, at least with the approach here adopted, when  $s$  is not constant, one requires statistics of order  $2r$  to obtain efficient estimates of the  $r$ -th moment or cumulant. Formally this involves an infinite regress. It may be that the problem will be soluble in finite terms by Robertson's or related methods.



The example shows that second order statistics may suffice for practical purposes when all values of  $s$  are of the order of 100 and not very variable. On the other hand had the same number of individuals occurred in some hundred samples in which  $s$  ranged from 1 to about 10, as in human families, fourth order statistics would have been desirable to obtain the correct weightings in evaluating  $k_2$ .

The numerical example shows that most of the published data on linkage are probably inaccurate. The mean recombination values found may only require slight revision. Their sampling errors are consistently larger than those published. The variances of recombination values will require estimation. A sufficiently variable value leads to spurious "interference" of the frequencies of crossing over in adjacent segments. In fact the whole theory of linkage will require revision when sufficient data are available. An attempt is being made to collect such data in this Institute.

I believe that the approach here outlined may be of a certain value in the design of sample surveys. We have seen that in the very simple case here considered, it is desirable, when  $k_2$  is large, to divide up the total population sampled into a large number of small samples even if the total cost or effort is thereby increased. This diminishes the sampling variance of the estimate  $k_1$ . However for a given total effort or cost the optimal design is not known till we have at least a rough estimate of  $\kappa_2$ . So ideally the procedure would be sequential. Further the optimal design for the estimation of  $\kappa_2$  is quite different from that optimal for the estimation of  $\kappa_1$ .

Sample surveys are seldom matters of mere counting. So the results of this investigation have no immediate relevance to them. But analogous problems will arise in sample surveys when variances as well as means are to be estimated not merely to determine the precision of means or differences between them, but for their own sake.

Some of the expressions here found can be derived, as limiting cases, from the theory of the analysis of variance. For example (1) to (6) can be derived by considering  $n$  sets of  $s$  samples, each of one member, the value of  $p_i$  being constant in each set. However I have not been able to obtain all the results of the paper by discussing such limiting cases, and the more direct approach here used can be applied to the evaluation of higher moments.

I have to thank Mrs Trent (Miss J. M. Clarke) for kindly putting her unpublished data at my disposal.

#### REFERENCES

- HALDANE, J. B. S. (1955): The rapid calculation of  $\chi^2$  as a test of homogeneity from a  $2 \times n$  table. *Biometrika*, **42**, 519-520.
- LEXIS, W. (1877): *Zur Theorie der Massenerscheinungen in der menschlichen Gesellschaft*, Freiburg.
- ROBERTSON, A. (1951): The analysis of heterogeneity in the binomial distribution. *Ann. Eugen.*, **16**, 1-14.

*Paper received : February, 1959.*



# SUFFICIENT STATISTICS OF MINIMAL DIMENSION<sup>1</sup>

By EDWARD W. BARANKIN<sup>2</sup>

and

MELVIN KATZ, Jr.<sup>3</sup>

*University of California, Berkeley*

**SUMMARY.** For families of probability distributions characterized by certain differentiability properties — a type of family customarily met in practice — the general problem of finding the smallest number of continuously differentiable, real-valued functions which can constitute a sufficient statistic is attacked. The precise local and global phases of this problem are formulated, and definitive results are obtained.

## 1. INTRODUCTION

Let  $\mathcal{P} = \{\mu_\theta, \theta \in \Theta\}$  be a family of probability measures on the Lebesgue-measurable subsets of an open set  $\Omega$  in  $E_0^n$ , a Euclidean  $n$ -space. A point of  $\Omega$  will be denoted by  $x = (x_1, x_2, \dots, x_n)$ . The parameter set  $\Theta$  will be an open subset of a Euclidean  $v$ -space,  $E_0^v$ ; thus,  $\theta = (\theta_1, \theta_2, \dots, \theta_v)$ . We suppose that each  $\mu_\theta$  is absolutely continuous with respect to Lebesgue measure, and we consider that there is a determination,  $p$ , of the family density function such that

$$p(x, \theta) > 0, \quad (x, \theta) \in \Omega \times \Theta, \quad \dots \quad (1.1)$$

and such that the following derivatives exist and are, together with  $p$ , continuous throughout  $\Omega \times \Theta$ :

$$\left. \begin{aligned} \frac{\partial p}{\partial x_i}, \quad i = 1, 2, \dots, n, \\ \frac{\partial p}{\partial \theta_j}, \quad j = 1, 2, \dots, v, \\ \frac{\partial^2 p}{\partial \theta_j \partial x_i}, \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, v, \\ \frac{\partial^2 p}{\partial x_i \partial \theta_j}, \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, v. \end{aligned} \right\} \quad \dots \quad (1.2)$$

<sup>1</sup> This paper was prepared with the partial support of the Office of Naval Research (Nonr-222-43). This paper in whole or in part may be reproduced for any purpose of the United States Government. Also this paper was supported (in part) by funds provided under Contract AF41(657)-29 with the School of Aviation Medicine, USAF, Randolph Air Force Base, Texas.

<sup>2</sup> During 1956-57, Fellow of the John Simon Guggenheim Memorial Foundation and Visiting Professor at the Institut Henri Poincaré, Paris, where a small portion of the research leading to this article was done. This author expresses his gratitude also to the Centre Universitaire International in Paris for the working facilities they so kindly provided.

<sup>3</sup> Now at the University of Chicago.



The continuity of the second derivatives implies that

$$\frac{\partial^2 p}{\partial \theta_j \partial x_i} = \frac{\partial^2 p}{\partial x_i \partial \theta_j}, \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, v. \quad \dots \quad (1.3)$$

The assumption of openness of  $\Omega$  and  $\Theta$  is not absolutely essential for the validity of our results, but we adopt it in order to avoid additional detail. In the kinds of cases that usually arise in practice, and in which  $\Omega$  and  $\Theta$  are not both open, it is an easy matter to show that our results, when applied to the interiors of  $\Omega$  and  $\Theta$ —which are open sets—give the desired results for  $\Omega \times \Theta$  itself.

The assumption (1.1) is to the effect that  $\Omega$  is the common carrier of all the densities of the family. Concerning the important class of families in which the carriers vary with  $\theta$ , we are hopeful that by suitable transformations of the problem it may be possible to apply bodily the theorems in this article—rather than re-prove them—in order to obtain the corresponding results.

The motivating question to which we address ourselves in this article is formulated and investigated in terms of the following two definitions.

*Definition 1.1:* A function  $T$  on  $\Omega$  will be said to be Euclidean of dimension  $r$  at  $x^0$  if there is a neighbourhood of  $x^0$  such that  $T$  maps this neighbourhood into a Euclidean  $r$ -space.

*Definition 1.2:* Let  $T$  be Euclidean of dimension  $r$  at  $x^0 \in \Omega$ ; specifically, let

$$T(x) = (h_1(x), h_2(x), \dots, h_r(x)) \quad \dots \quad (1.4)$$

in some neighbourhood of  $x^0$ . Then we say that  $T$  is regular at  $x^0$  if the functions  $h_i$ ,  $i = 1, 2, \dots, r$  are continuously differentiable and if the Jacobian matrix

$$J(x) = \left\| \frac{\partial h_i}{\partial x_j} \right\|_{\substack{i=1, 2, \dots, r \\ j=1, 2, \dots, n}} \quad \dots \quad (1.5)$$

is of rank  $r$  at  $x^0$ .

Notice that in this last definition, by virtue of the continuous differentiability of the  $h_i$ , the matrix  $J$  is of rank  $r$  everywhere in a neighbourhood of  $x^0$ . Notice also that this rank condition implies  $r \leq n$ .

We may now pose this question: for a given point  $x^0 \in \Omega$ , what is the smallest integer  $r$  such that there exists a sufficient statistic for  $\mathcal{P}$  which is Euclidean of dimension  $r$  at  $x^0$ , and continuously differentiable in a neighbourhood of  $x^0$ ? Under the limited conditions we have imposed on the family density function  $p$ , there is no reason to expect that this smallest integer  $r$  will be the same for every point  $x^0$ . And indeed, our second example in Section 5 shows that this *local* minimal dimension



## SUFFICIENT STATISTICS OF MINIMAL DIMENSION

of a sufficient statistic may vary over  $\Omega$ . The answer to the above question, for points  $x^0$  of  $\Omega$  that we call *regular points*, is given by Theorems 3.2 and 3.3. More exactly, the sequence of results is as follows. Definitions 3.1 and 3.2 present a certain explicit, integer-valued function  $\rho$  on  $\Omega$ . Then Theorem 3.1 shows that for any point  $x^0 \in \Omega$ ,  $\rho(x^0)$  is a lower bound for the dimension at  $x^0$  of a sufficient statistic which is Euclidean at  $x^0$  and regular at  $x^0$ . By means of an additional argument, we give a direct strengthening of this result in proving, in Theorem 3.2, that  $\rho(x^0)$  is a lower bound for the dimension at  $x^0$  of a sufficient statistic which is Euclidean at  $x^0$  and continuously differentiable in a neighbourhood of  $x^0$  (but not necessarily such that the matrix  $J(x^0)$  is of the pertinent rank  $r$ ). Following this, Definition 3.3 characterizes a regular point of  $\Omega$ , and thereupon we show (after proving two lemmas), with Theorem 3.3, that at a regular point  $x^0$  the lower bound  $\rho(x^0)$  is achieved. This theorem is proved constructively, so that it gives an explicit sufficient statistic which is dimensionally minimal at  $x^0$ . These results comprise Section 3. Section 2 is devoted to establishing a series of lemmas toward the proof of Theorem 3.1.

We go on in Section 4 to obtain global results out of the local results of Section 3. Evidently, the extent of globality that we may hereby obtain is necessarily limited to the set  $R$  of regular points of  $\Omega$ . This set is an open set, and its complement in  $\Omega$ —the set of nonregular points—is nowhere dense in  $\Omega$  (see Lemma 3.1). In general, the Lebesgue measure of the set of nonregular points need not be 0. However, to encounter a case in which this nowhere dense set is of positive measure would be the unusual thing. In all cases of practical importance this set is of measure 0. Thus, in the stochastic context, our results in Section 4, being valid almost everywhere (Lebesgue) in  $R$ , are, for all practical purposes, fully global for  $\Omega$ . Theorem 4.1 establishes (constructively) the basic global result that there is a sufficient statistic,  $T^*$ , which is of minimal local dimension and regular almost everywhere in  $R$ .  $T^*$  is continuously differentiable on an open subset of  $R$  whose complement in  $R$  is of Lebesgue measure 0; and among all sufficient statistics having this desirable analytical property, such a statistic as  $T^*$  is the ultimate answer to the question of a minimal dimensionality, as far as  $R$  is concerned—and as far as  $\Omega$  is concerned, if  $\Omega - R$  is of measure 0.

The statistic  $T^*$  has, in general, different dimensions at various points of its continuously differentiable domain. One may prefer to deal rather with an almost everywhere continuously differentiable sufficient statistic of *constant* dimension over its continuously differentiable domain, and therefore desire to know the minimum possible value for this *global* dimension. This is in the spirit of most usage in the literature. Theorem 4.2 contributes to the answer to this question. Of course, the minimal global dimension achievable within  $R$  is  $\max_{x \in R} \rho(x)$ .

Finally in Section 4, we raise the obviously pertinent question of how far dimensional minimality of a sufficient statistic carries us toward functional minimality. By a (globally) *functionally minimal* sufficient statistic we mean one which is a function



(almost everywhere relative to each of the measures in  $\mathcal{P}$ ) of any other sufficient statistic. This is the minimality notion discussed by Lehmann and Scheffé (1950). Their result, Theorem 6.3, p. 336, shows that for the family  $\mathcal{P}$  that we are concerned with here there does exist a functionally minimal sufficient statistic. We are able to show here, in Theorem 4.3, that the sufficient statistic  $T^*$  of Theorem 4.1, or, equally well, the statistic  $T^+$  of Theorem 4.2., is *locally* functionally minimal almost everywhere in  $R$ . That is, for almost all points of  $R$ , there is a neighbourhood of a point such that, in this neighbourhood,  $T^*$ , or  $T^+$ , is a function of any sufficient statistic. Whether or not this can be improved to global functional minimality almost everywhere in  $R$ , while preserving the desired property of continuous differentiability, is left an open question.

Section 5 is the last section of the paper. In it we present two examples to illustrate our results.

For our investigation we shall need a general criterion for a sufficient statistic for our family  $\mathcal{P}$ . This is provided, in most convenient form, by a result of Bahadur (1954, Corollary 6.1, p. 438); this result has evolved from Fisher (1922) and the theorem of Neyman (1935) through the work of Halmos and Savage (1949, Corollary 1, p. 234) and Lehmann and Scheffé (1950, Theorem 6.2, p. 332). For easy reference, we shall state Bahadur's criterion here as a lemma, in terms specific to our particular case. The statistic  $T$  appearing in the statement has no prior conditions on it; it is any particular function on  $\Omega$ , and it is rendered measurable by taking, as the pertinent  $\sigma$ -algebra of subsets of its range, the class of all sets whose inverse images, under  $T$ , are Lebesgue sets. If  $\mathcal{R}_T$  denotes the range of  $T$ , the criterion is as follows.

**Lemma 1.1:** (*N.—H.—Sa.—L.—Sc.—B*). *A necessary and sufficient condition that the statistic  $T$  be a sufficient statistic for  $\mathcal{P}$  is that there exists a nonnegative function  $f$  on  $\mathcal{R}_T \times \Theta$ , and a nonnegative function  $g$  on  $\Omega$ , such that (i) for each  $\theta \in \Theta$ ,  $f(T(\cdot), \theta)$  is Lebesgue measurable, (ii)  $g$  is Lebesgue measurable, and (iii) for each  $\theta \in \Theta$  the equality*

$$p(x, \theta) = f(T(x), \theta)g(x) \quad \dots \quad (1.6)$$

*holds for almost all (Lebesgue)  $x \in \Omega$ .*

We shall make abundant use of the implicit function theorem. For an account of this the reader is referred to, for example, Caratheodory (1935, p. 9). The other analytical tools we employ, such as transformation of integrals, the Fubini theorem, etc., need no special references.

Results of the kind achieved in this article are important for investigations wherein one wants to have the advantage of differentiable statistics. Prominent among authors whose work is concerned with such sufficient statistics are Bhattacharyya and Rao (see References). In addition, the question arises as to the possibility of extending the now classical results of Fisher (1934) and Darmois and Koopman (see References) with the aid of results of the kind we obtain here. We shall not consider these applications in this article.



## SUFFICIENT STATISTICS OF MINIMAL DIMENSION

The article of Dynkin (1951)—which has not appeared in the references of any of the English language articles on sufficient statistics, and of which the authors became aware only when the present paper was in its final stages of preparation—is of interest on two accounts relative to our work here. First, Dynkin pursues his investigation of sufficient statistics with attention to their local functional character; secondly, his Theorem 2 is an instance of a result determining, in a special case (product distributions), a sufficient statistic which is, locally, dimensionally and functionally minimal. In regard to the first point, the authors thus find another, and apparently the only other, investigator who brings into evidence the essential role of local considerations in regard to minimal dimensionality. In the present article, the general local-versus-global situation is laid open rather completely.

We remark that the problem of minimal dimensionality of sufficient statistics was originally proposed to one of the authors by J. Neyman.

### 2. THREE LEMMAS

The goal of this section is the establishment of the result stated as Lemma 2.3 below. This is the important preliminary result that if the sufficient statistic  $T$  is Euclidean at  $x^0$  and regular at  $x^0$  then the factorization (1.6) can be made locally with fully differentiable functions  $f$  and  $g$ . Lemmas 2.1 and 2.2 are ancillary to the proof of Lemma 2.3.

**Lemma 2.1:** *Let the function  $T$  on  $\Omega$  be Euclidean of dimension  $r$  at  $x^0$  and regular at  $x^0$ , being given by (1.4) in some neighbourhood,  $N$ , of  $x^0$ . Let  $x'$  and  $x''$  be two points of  $N$  such that  $J(x')$  and  $J(x'')$  (see (1.5)) are both of rank  $r$ , and such that  $T(x') = T(x'')$ .*

*Let  $N'$  and  $N''$  be disjoint neighbourhoods of  $x'$  and  $x''$ , respectively, and let  $A \subseteq N'$  and  $A'' \subseteq N''$  be Lebesgue measurable sets such that the sets  $N' - A'$  and  $N'' - A''$  are of Lebesgue measure 0.*

*Then, there exist points  $x^I \in A'$  and  $x^{II} \in A''$ , different from  $x'$  and  $x''$ , respectively, such that  $T(x^I) = T(x^{II})$ .*

The proof of this lemma involves more complication for  $r < n$  than for  $r = n$ . We shall carry through the arguments for the case  $r < n$ , and indicate at the proper place the simplification for  $r = n$ .

*Proof:* Consider the point  $x'$ . We may suppose, without loss of generality, that

$$\left( \frac{\partial(h_1, h_2, \dots, h_r)}{\partial(x_1, x_2, \dots, x_r)} \right)_{x'} \neq 0. \quad \dots \quad (2.1)$$

Let  $E^r$  denote the  $r$ -dimensional Euclidean range-space of  $T$  on  $N$ , and let  $y$  denote a point of  $E^r$ . Let  $E_0^n$  (the containing space of  $\Omega$ ) be represented as  $E_0^r \times E_0^{n-r}$ , so that for any point  $x = (x_1, x_2, \dots, x_n) \in E_0^n$ , we have  $(x_{r+1}, x_{r+2}, \dots, x_n) \in E_0^{n-r}$ . A point of  $E_0^{n-r}$  will be denoted by  $z$ ; and  $z'$  will denote the specific point  $(x'_{r+1}, x'_{r+2}, \dots, x'_n)$ —that is, the projection of  $x'$  into  $E_0^{n-r}$ . Finally, let  $U$  denote the identity on  $E_0^{n-r}$ .



It follows from (2.1), by the Implicit Function Theorem, that there is a spherical neighbourhood  $S'$  in  $E^r$ , centered at  $T(x')$ , and a spherical neighbourhood  $W'$  in  $E_0^{n-r}$ , centered at  $z'$ , and a continuously differentiable (vector) function  $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_r)$  on  $S' \times W'$  into  $E_0^r$ , such that the continuously differentiable function  $(\varphi, U)$  on  $S' \times W'$  is onto a neighbourhood,  $D'$ , of  $x'$ , and is the inverse of the continuously differentiable function  $(T, U)$  on  $D'$  onto  $S' \times W'$ . We take  $S'$  and  $W'$  so small that (i)  $D' \subseteq N'$ , (ii)  $D'$  is bounded, and (iii)

$$\left( \frac{\partial(h_1, h_2, \dots, h_r)}{\partial(x_1, x_2, \dots, x_r)} \right)_{x \in D'} \neq 0. \quad \dots (2.2)$$

We have then

$$\frac{\partial(\varphi_1, \varphi_2, \dots, \varphi_r, x_{r+1}, \dots, x_n)}{\partial(y_1, y_2, \dots, y_r, x_{r+1}, \dots, x_n)} = \frac{\partial(\varphi_1, \varphi_2, \dots, \varphi_r)}{\partial(y_1, y_2, \dots, y_r)} = \frac{1}{\left( \frac{\partial(h_1, h_2, \dots, h_r)}{\partial(x_1, x_2, \dots, x_r)} \right)_{\substack{x=(\varphi(y, z), z) \\ (y, z) \in S' \times W'}}}}. \quad \dots (2.3)$$

The set  $N' - A'$  is a subset of  $N'$  of Lebesgue measure 0. It is a Lebesgue-measurable set, but not necessarily a Borel set. We may, however, choose a Borel set in  $N'$  of Lebesgue measure 0 which includes  $N' - A'$ ; let  $B'_1$  be such a set. Then  $B' = N' - B'_1$  is likewise a Borel set and is a subset of  $A'$ . Thus, we have replaced  $A'$  by a Borel set  $B' \subseteq A'$  which also has the property that  $N' - B'$  is of Lebesgue measure 0.

Let  $C' = B' \cap D'$ . Since  $D'$  is an open set,  $C'$  is a Borel set, and  $D' - C'$  is of Lebesgue measure 0. Hence, if  $\lambda_n$  denotes  $n$ -dimensional Lebesgue measure, we have

$$\lambda_n(C') = \lambda_n(D'). \quad \dots (2.4)$$

This common measure is finite since  $D'$  is bounded.

$D' \subseteq E_0^n$  being the image of  $S' \times W' \subseteq E^r \times E_0^{n-r}$  under  $(\varphi, U)$ , we have

$$\lambda_n(D') = \int_{S' \times W'} \left| \frac{\partial(\varphi_1, \varphi_2, \dots, \varphi_r, x_{r+1}, \dots, x_n)}{\partial(y_1, y_2, \dots, y_r, x_{r+1}, \dots, x_n)} \right| d\lambda_n. \quad \dots (2.5)$$

For the sake of brevity, let  $M(y, z)$  denote the integrand values of (2.5), that is, the common absolute value of the three members of (2.3). Thus, (2.5) becomes

$$\lambda_n(D') = \int_{S' \times W'} M d\lambda_n. \quad \dots (2.6)$$

Let  $K'$  denote the image of  $C'$  under  $(T, U)$ . Then  $K'$  is the inverse image of  $C'$  under  $(\varphi, U)$ , and since  $(\varphi, U)$  is continuous and  $C'$  is a Borel set, it follows that  $K'$  is a Borel set. Consequently, we have

$$\lambda_n(C') = \int_{K'} M d\lambda_n. \quad \dots (2.7)$$



If we define the function  $k$  on  $S' \times W'$  by

$$k(y, z) = \begin{cases} 1 & \text{if } (y, z) \in K', \\ 0 & \text{otherwise,} \end{cases} \quad \dots \quad (2.8)$$

then (2.7) may be replaced by

$$\lambda_n(C') = \int_{S' \times W'} k M d\lambda_n. \quad \dots \quad (2.9)$$

Let  $\lambda_r$  denote  $r$ -dimensional Lebesgue measure in the  $r$ -dimensional sphere  $S'$ , and  $\lambda_{n-r}$  denote  $(n-r)$ -dimensional Lebesgue measure in the  $(n-r)$ -dimensional sphere  $W'$ . We apply the Fubini theorem to (2.6) and (2.9) to obtain

$$\lambda_n(D') = \int_{W'_1} \left[ \int_{S'} M d\lambda_r \right] d\lambda_{n-r} \quad \dots \quad (2.10)$$

and

$$\lambda_n(C') = \int_{W'_1} \left[ \int_{S'} k M d\lambda_r \right] d\lambda_{n-r}, \quad \dots \quad (2.11)$$

where  $W'_1$  is a subset of  $W'$ , with  $\lambda_{n-r}(W'_1) = \lambda_{n-r}(W')$ , such that for each  $z \in W'_1$ , the integrands in (2.10) and (2.11) are  $\lambda_r$ -measurable and integrable functions (of  $y$ ) on  $S'$ . Now, the integrand in (2.11) is everywhere less than or equal to the integrand in (2.10). Therefore, for each  $z \in W'_1$ , the inner integral in (2.11) is less than or equal to the inner integral in (2.10). Hence, by (2.4) it follows that for almost all  $z \in W'_1$ , the inner integral in (2.11) is equal to the inner integral in (2.10). Let  $z^I$  be such a point of  $W'_1$ ; thus,

$$\int_{S'} k(., z^I) M(., z^I) d\lambda_r = \int_{S'} M(., z^I) d\lambda_r. \quad \dots \quad (2.12)$$

The function  $M(., z^I)$  on  $S'$  is positive throughout  $S'$ , and therefore (2.12) implies that

$$k(y, z^I) = 1 \quad \text{for almost all } y \in S'. \quad \dots \quad (2.13)$$

Now, for a given  $y$ , there is a  $z$  such that  $k(y, z) = 1$  if and only if there is a  $z$  such that  $(y, z) \in K'$ , thus, if and only if there is an  $x \in C'$  such that  $T(x) = y$ . Hence, (2.13) asserts that *for almost all  $y \in S'$  there is a point  $x \in C'$  such that  $T(x) = y$* . Recalling that  $C' \subseteq B' \subseteq A'$ , we may restate this as : *for almost all  $y \in S'$  there is a point  $x \in A'$  such that  $T(x) = y$* .

This is the result we have sought concerning the set  $A'$ . As we indicated we would do, we have given the explicit arguments for the case  $r < n$ . However, it is now evident that in the case  $r = n$  the arguments proceed in the same way, except that the factor space  $E_0^{n-r}$ , and the sphere  $W'$  in it, do not enter into the considerations, and it is therefore no longer necessary to appeal to the Fubini theorem.



We now carry out the same kind of analysis for the set  $A''$ , and we obtain the corresponding result, namely, that for almost all  $y$  in some sphere  $S'' \subseteq E^r$ , centered at  $T(x'') = T(x')$ , there is a point  $x \in A''$  such that  $T(x) = y$ . Thus, the spheres  $S'$  and  $S''$  (are in the same space,  $E^r$ , and) are concentric, and if  $S$  denotes the smaller of the two, then our results on  $A'$  and  $A''$  combine to give the result that, for almost all  $y \in S$ , there are points  $x^I \in A'$  and  $x^{II} \in A''$  such that  $T(x^I) = T(x^{II}) = y$ . Hence, on choosing such a  $y \neq$  the center of  $S$ , we have the result asserted in the statement of Lemma 2.1. The proof is therefore complete.

Lemma 2.2: *Under the hypothesis of Lemma 2.1, there exist sequences  $\{x'^{(s)}, s = 1, 2, \dots\} \subseteq A'$  and  $\{x''^{(s)}, s = 1, 2, \dots\} \subseteq A''$ , tending properly (that is, no point of the sequence is identical with the limit point) to  $x'$  and  $x''$ , respectively, such that  $T(x'^{(s)}) = T(x''^{(s)})$  for all  $s = 1, 2, \dots$ .*

This is an immediate consequence of Lemma 2.1.

Lemma 2.3: *Let  $T$  be a  $\mathcal{P}$ -sufficient statistic which is Euclidean of dimension  $r$  at  $x^0$  and regular at  $x^0$ , being given by (1.4) in some neighbourhood of  $x^0$ .*

*Then, there is a neighbourhood  $N$  of  $x^0$  such that  $T(N)$  is a neighbourhood of  $T(x^0)$  and there are functions  $f$ , on  $T(N) \times \Theta$ ; and  $g$ , on  $N$ , such that*

$$p(x, \theta) = f(T(x), \theta) g(x) \text{ for all } x \in N, \theta \in \Theta, \quad \dots (2.14)$$

*and such that  $f(y, \theta)$  has continuous partial derivatives  $\frac{\partial f}{\partial y_i}, \frac{\partial f}{\partial \theta_j}, \frac{\partial^2 f}{\partial y_i \partial \theta_j} = \frac{\partial^2 f}{\partial \theta_j \partial y_i}$ ,  $i = 1, 2, \dots, r; j = 1, 2, \dots, v$ , in  $T(N) \times \Theta$ , and  $g(x)$  has continuous partial derivatives  $\frac{\partial g}{\partial x_i}, i = 1, 2, \dots, n$ , in  $N$ .*

*Proof:* Let  $N_0$  be a neighbourhood of  $x^0$  such that  $J(x)$  is of rank  $r$  for all  $x \in N_0$ . It is a consequence of the Implicit Function Theorem that the image under  $T$  of any neighbourhood of  $x^0$  includes a neighbourhood of  $T(x^0)$  in  $E^r$ . Let  $N^{(r)}$  be a neighbourhood of  $T(x^0)$  which is included in  $T(N_0)$ . And define

$$N = N_0 \cap T^{-1}(N^{(r)}). \quad \dots (2.15)$$

It follows from the continuity of  $T$  that  $N$  is an open set, and hence a neighbourhood of  $x^0$ .

If  $x \in N$ , then  $x \in T^{-1}(N^{(r)})$ , and therefore  $T(x) \in N^{(r)}$ . Conversely, suppose  $y$  is any particular point of  $N^{(r)}$ . Since  $N^{(r)} \subseteq T(N_0)$ , there is a point  $x \in N_0$  such that  $y = T(x)$ . For this  $x$  we have both  $x \in N_0$  and  $x \in T^{-1}(N^{(r)})$ ; therefore,  $x \in N$ . We have thus proved that

$$T(N) = N^{(r)}, \quad \dots (2.16)$$

which is to say that for the neighbourhood  $N$  of  $x^0$ ,  $T(N)$  is a neighbourhood of  $T(x^0)$  in  $E^r$ .



# SUFFICIENT STATISTICS OF MINIMAL DIMENSION

Let us keep in mind also that, since  $N \subseteq N_0$ , the matrix  $J(x)$  is of rank  $r$  for all  $x \in N$ .

Since  $T$  is sufficient, we have, by Lemma 1.1, that there exist nonnegative functions  $f_1$ , on  $\mathcal{R}_T \times \Theta$ , and  $g_1$ , on  $\Omega$ , such that, for each  $\theta \in \Theta$ ,

$$p(x, \theta) = f_1(T(x), \theta) g_1(x) \quad \dots \quad (2.17)$$

for almost all (Lebesgue)  $x \in \Omega$ . We shall confine our attention to this equation for  $x \in N$ . For any  $\theta$  we have  $p(x, \theta) > 0$  for all  $x \in N$  (see (1.1)), and therefore (2.17) implies that  $f_1(T(x), \theta) > 0$  and  $g_1(x) > 0$  for almost all  $x \in N$ . By virtue of this, if we now choose a particular point  $\theta^0 \in \Theta$ , which will be held fixed in its role for the rest of this proof, we then obtain from (2.17) that

$$\frac{p(x, \theta)}{p(x, \theta^0)} = \frac{f_1(T(x), \theta)}{f_1(T(x), \theta^0)}, \text{ a.e. } x \in N, \theta \in \Theta. \quad \dots \quad (2.18)$$

Thus, for each  $\theta$ ,  $\frac{p(x, \theta)}{p(x, \theta^0)}$  is a function of  $T(x)$  almost everywhere in  $N$ . We shall now show that, in fact,  $\frac{p(x, \theta)}{p(x, \theta^0)}$  is a function of  $T(x)$  *everywhere* in  $N$ .

For a given  $\theta$ , let  $A$  be the subset of  $N$  for each point of which (2.18) holds. Then  $N - A$  is of Lebesgue measure 0. Let  $x'$  be a point of  $N - A$ . Suppose  $x''$  is another point of  $N$  such that  $T(x'') = T(x')$ . We wish to show that the left-hand side of (2.18) has the same value for  $x'$  and  $x''$ . We apply Lemma 2.2 to obtain sequences  $\{x'^{(s)}\} \subseteq A$  and  $\{x''^{(s)}\} \subseteq A$ , converging properly to  $x'$  and  $x''$ , respectively, and such that  $T(x'^{(s)}) = T(x''^{(s)})$  for all  $s = 1, 2, \dots$ . It follows then, from (2.18), that

$$\frac{p(x''^{(s)}, \theta)}{p(x''^{(s)}, \theta^0)} = \frac{p(x'^{(s)}, \theta)}{p(x'^{(s)}, \theta^0)}, \quad s = 1, 2, \dots \quad \dots \quad (2.19)$$

From this it follows immediately, by virtue of the continuity of  $p$ , that

$$\frac{p(x'', \theta)}{p(x'', \theta^0)} = \frac{p(x', \theta)}{p(x', \theta^0)}. \quad \dots \quad (2.20)$$

And this is what was to be established.

Thus,  $\frac{p(x, \theta)}{p(x, \theta^0)}$  is a function of  $T(x)$  everywhere in  $N$ ; and this is true for each  $\theta \in \Theta$ . Hence, there is a function  $f$ , on  $T(N) \times \Theta$ , such that

$$\frac{p(x, \theta)}{p(x, \theta^0)} = f(T(x), \theta) \quad x \in N, \theta \in \Theta. \quad \dots \quad (2.21)$$



If we define the function  $g$  on  $N$  by

$$g(x) = p(x, \theta^0), \quad x \in N, \quad \dots \quad (2.22)$$

then we have

$$p(x, \theta) = f(T(x), \theta) g(x), \quad x \in N, \theta \in \Theta. \quad \dots \quad (2.23)$$

It remains now to show that  $f$  and  $g$  have the continuous differentiability properties asserted by Lemma 2.3. This result for the function  $g$  is immediate from (2.22), by virtue of the differentiability of  $p$ . We turn our attention to the function  $f$ .

Let  $y' = (y'_1, y'_2, \dots, y'_r)$  be any particular point of  $T(N)$ . Let  $x' = (x'_1, x'_2, \dots, x'_n)$  be a point of  $N$  such that  $T(x') = y'$ . The matrix  $J(x)$  is of rank  $r$  at  $x'$ ; let us suppose, for simplicity, that

$$\left( \frac{\partial(h_1, h_2, \dots, h_r)}{\partial(x_1, x_2, \dots, x_r)} \right)_{x'} \neq 0. \quad \dots \quad (2.24)$$

We now have the situation that we encountered in the proof of Lemma 2.1, and to which we applied the Implicit Function Theorem to obtain the continuously differentiable inverse,  $(\varphi, U)$  of  $(T, U)$ . The function  $(\varphi, U)$  is defined on  $S' \times W'$ , where  $S'$  is a spherical neighbourhood in  $E^r$ , centered at  $y'$ , and  $W'$  is a spherical neighbourhood in  $E^{n-r}$ , centered at  $z' = (x'_{r+1}, x'_{r+2}, \dots, x'_n)$ . (Again as in the proof of Lemma 2.1, there is no need to consider a  $W'$  in the case  $r = n$ .) For the present context,  $S'$  and  $W'$  may be considered chosen small enough so that  $D' \subseteq N$ . Then (2.21) gives us the following :

$$f(y, \theta) = \frac{p(\varphi(y, z), z, \theta)}{p(\varphi(y, z), z, \theta^0)}, \quad y \in S', z \in W', \theta \in \Theta. \quad \dots \quad (2.25)$$

From this expression for  $f(y, \theta)$  in  $S' \times \Theta$ , because of the continuous differentiability of  $p$  and  $\varphi$ , we get immediately the asserted continuous differentiability of  $f$  in  $S' \times \Theta$ . Since this result obtains for every particular point  $y' \in T(N)$ , we have established the asserted continuous differentiability of  $f$  in  $T(N) \times \Theta$ .

Lemma 2.3 is now fully established.

### 3. THE LOCAL MINIMAL DIMENSION OF A SUFFICIENT STATISTIC

The three theorems of this section constitute the basic results of this article; it is from them that we are able to deduce global results in the next section. We refer the reader to the Introduction for a descriptive outline of the sequence of conclusions toward which we now begin to argue.

For any particular point  $x \in \Omega$ , let  $\{j_1, j_2, \dots, j_n\}$  be any particular collection of  $n$  integers, each chosen from the set  $\{1, 2, \dots, v\}$ , and let  $\{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n)}\}$  be any



# SUFFICIENT STATISTICS OF MINIMAL DIMENSION

particular collection of points of  $\Theta$  (not necessarily all distinct). We define the  $n \times n$  matrix

$$L(x; j_1, j_2, \dots, j_n; \theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n)}) = \left\| \left( \frac{\partial^2 \log p}{\partial x_i \partial \theta_{j_k}} \right)_{x, \theta^{(k)}} \right\|_{i, k=1, 2, \dots, n}. \quad \dots (3.1)$$

Let  $\mathcal{L}_x$  denote the class of all matrices (3.1) for the given point  $x$ .

*Definition 3.1:* The integer-valued function  $\rho_1$ , on  $\Omega$ , is defined by

$$\rho_1(x) = \max_{L \in \mathcal{L}_x} (\text{rank } L). \quad \dots (3.2)$$

The symbol  $\mathcal{S}_{x^0, \alpha}$  will denote the open sphere in  $\Omega$  centered at  $x^0$  and of radius  $\alpha$ .

*Definition 3.2:* The integer-valued function  $\rho$ , on  $\Omega$ , is defined by

$$\rho(x^0) = \lim_{\alpha \rightarrow 0} \max_{x \in \mathcal{S}_{x^0, \alpha}} \rho_1(x). \quad \dots (3.3)$$

With these definitions we now prove

*Theorem 3.1:* If  $T$  is a  $\mathcal{P}$ -sufficient statistic which is Euclidean of dimension  $r$  at  $x^0$  and regular at  $x^0$ , then

$$r \geq \rho(x^0). \quad \dots (3.4)$$

*Proof:* Let  $T$  be given by (1.4) in a neighbourhood  $x^0$ . By Lemma 2.3, there is a neighbourhood  $N$  of  $x^0$ , and differentiable functions  $f$  and  $g$ , as detailed in the Lemma, such that

$$p(x, \theta) = f(T(x), \theta) g(x), \quad x \in N, \quad \theta \in \Theta. \quad \dots (3.5)$$

We may differentiate this to obtain

$$\frac{\partial^2 \log p}{\partial x_i \partial \theta_j} = \sum_{s=1}^r \left( \frac{\partial^2 \log f}{\partial y_s \partial \theta_j} \right)_{y=T(x)} \left( \frac{\partial h_s}{\partial x_i} \right), \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, v. \quad (3.6)$$

It follows that for any particular  $x \in N$ , and a set of integers  $\{j_1, j_2, \dots, j_n\}$  and a set of points  $\{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n)}\}$  which define a matrix (3.1) for this  $x$ , we have the factorization

$$L(x; j_1, j_2, \dots, j_n; \theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n)}) = J(x) \cdot \left\| \left( \frac{\partial^2 \log f}{\partial y_s \partial \theta_{j_k}} \right)_{T(x), \theta^{(k)}} \right\|_{\substack{s=1, 2, \dots, r \\ k=1, 2, \dots, n}} \quad \dots (3.7)$$

The rank of the product of two matrices is not greater than the rank of either factor; the first factor on the right-hand side of (3.7) has rank  $r$  (and the second factor has rank at most  $r$ ). It follows that the rank of  $L$  in (3.7) is  $\leq r$ . Since this is true for every  $L \in \mathcal{L}_x$ , we have  $\rho_1(x) \leq r$ .



This result holds for each  $x \in N$ . Therefore, if  $\alpha > 0$  is small enough so that  $S_{x^0, \alpha} \subseteq N$ , then we have  $\max_{x \in S_{x^0, \alpha}} \rho_1(x) \leq r$ . Since, clearly, for any sphere  $S_{x^0, \alpha}$ , it is true that  $\rho(x^0) \leq \max_{x \in S_{x^0, \alpha}} \rho_1(x)$ , we get, finally, that  $\rho(x^0) \leq r$ . This is the asserted result (3.4), and the theorem is therefore established.

Theorem 3.1 may now be strengthened, as follows.

Theorem 3.2: *If  $T$  is a  $\mathcal{P}$ -sufficient statistic which is Euclidean of dimension  $r$  at  $x^0$  and continuously differentiable in a neighbourhood of  $x^0$  (but not necessarily such that the matrix  $J(x^0)$  is of rank  $r$ ), then*

$$r \geq \rho(x^0). \quad \dots (3.8)$$

*Proof:* Since  $\rho_1$  is an integer-valued function (that is, its range is a discrete set), it follows that (1) for all  $x$  in a sufficiently small sphere centered at  $x^0$ , we have  $\rho_1(x) \leq \rho(x^0)$ , and (2) in every sphere centered at  $x^0$ , of however small radius, there is a point  $x'$  with  $\rho_1(x') = \rho(x^0)$ . The fact (1) then implies that for the points  $x'$  of (2) which lie in a sufficiently small sphere centered at  $x^0$ , we have  $\rho(x') = \rho_1(x') = \rho(x^0)$ . The continuity of the derivatives of  $p$  implies, finally, that for such a point  $x'$  sufficiently close to  $x^0$ , there is a neighbourhood of  $x'$  at every point  $x$  of which we have  $\rho(x) = \rho(x') = \rho(x^0)$ . These considerations show that in any neighbourhood of  $x^0$ , however small, there is an open subset, for every point  $x$  of which we have  $\rho(x) = \rho(x^0)$ . (See the proof of Lemma 3.1 below for complete details on these arguments.)

Let  $T$  be given by (1.4) in a neighbourhood  $N$  of  $x^0$ . Let  $N' \subseteq N$  be (according to the preceding paragraph) such that

$$\rho(x) = \rho(x^0), \quad x \in N'. \quad \dots (3.9)$$

Define 
$$r' = \max_{x \in N'} [\text{rank } J(x)], \quad \dots (3.10)$$

and let  $x'$  be a particular point of  $N'$  such that

$$\text{rank } J(x') = r'. \quad \dots (3.11)$$

We now have the situation that  $J(x)$  is of rank  $r'$  at  $x'$  (by (3.11)), and for each  $x \in N'$  (a neighbourhood of  $x'$ ), every minor of  $J(x)$  of order  $> r'$  vanishes (by (3.10)). It follows, by a familiar course of reasoning from the Implicit Function Theorem, that there is a neighbourhood  $N''$  (a subset of  $N'$ ) of  $x'$ , and some specific subset of  $r'$  of the functions  $h_1, h_2, \dots, h_r$ , such that the Jacobian matrix of these  $r'$  functions is of rank  $r'$  throughout  $N''$  and such that  $T$  is a continuously differentiable function of only these  $r'$  functions in  $N''$ . If the  $r'$  functions in question are  $h_{i_1}, h_{i_2}, \dots, h_{i_{r'}}$ , then define  $T'$  as follows :

$$T'(x) = \begin{cases} (h_{i_1}(x), h_{i_2}(x), \dots, h_{i_{r'}}(x)), & x \in N'', \\ T(x), & x \in \bar{\Omega} - N''. \end{cases} \quad \dots (3.12)$$



# SUFFICIENT STATISTICS OF MINIMAL DIMENSION

Clearly,  $T'$  is Euclidean of dimension  $r'$  at  $x'$  and regular at  $x'$ . Moreover, it is a ready consequence of Lemma 1.1 that, since  $T$  is a sufficient statistic, so also is  $T'$ . Hence, by Theorem 3.1,

$$r' \geq \rho(x'). \quad \dots (3.13)$$

Now, finally, we have  $r \geq r'$  and, by (3.9) since  $x' \in N'$ ,  $\rho(x') = \rho(x^0)$ . Applying (3.13) to these two facts we get (3.8), and the proof of Theorem 3.2 is complete.

Having thus established that the function  $\rho$  provides a lower bound for the local dimension of a locally continuously differentiable, Euclidean sufficient statistic, we turn now to the task of showing that this lower bound is attained at points of  $\Omega$  of a certain kind. The following definition characterizes the kind of point in question.

*Definition 3.3:* A point  $x \in \Omega$  is said to be a regular point of  $\Omega$  (for the family  $\mathcal{T}$ ) if

$$\rho(x) = \rho_1(x). \quad \dots (3.14)$$

In order to have an idea of the incidence of regular points in  $\Omega$ , before going on to derive results concerning them, we prove the following lemma.

*Lemma 3.1:* The set  $R$ , of regular points of  $\Omega$ , is an everywhere dense, open subset of  $\Omega$ . (Equivalently, the set  $\Omega - R$ , of nonregular points of  $\Omega$ , is a nowhere dense subset of  $\Omega$ , closed in  $\Omega$ .)

*The function  $\rho$  is continuous on  $R$ .*

*Proof:* Let  $N$  be any open subset of  $\Omega$ . We shall show that  $N$  contains a regular point, thereby proving that  $R$  is everywhere dense. Let  $x^0$  be a particular point of  $N$ . If every sphere  $\mathcal{S}_{x^0, \alpha}$  contained a point  $x$  with  $\rho_1(x) > r$  (some particular integer), then, by Definition 3.2, we should have  $\rho(x^0) \geq r + 1$ . In particular on taking  $r = \rho(x^0)$ , we should have the self-contradictory assertion that  $\rho(x^0) \geq \rho(x^0) + 1$ ; and therefore it follows that there is a sphere  $\mathcal{S}_{x^0, \alpha_1} \subseteq N$  such that

$$\rho_1(x) \leq \rho(x^0), \quad x \in \mathcal{S}_{x^0, \alpha_1}. \quad \dots (3.15)$$

If the strict inequality held in (3.15) for all  $x \in \mathcal{S}_{x^0, \alpha_1}$ , then it would likewise hold for all spheres  $\mathcal{S}_{x^0, \alpha}$  with  $\alpha < \alpha_1$ , and therefore we should have  $\max_{x \in \mathcal{S}_{x^0, \alpha}} \rho_1(x) \leq \rho(x^0) - 1$

for  $\alpha < \alpha_1$ , and hence  $\rho(x^0) = \lim_{\alpha \rightarrow 0} \max_{x \in \mathcal{S}_{x^0, \alpha}} \rho_1(x) \leq \rho(x^0) - 1$ , which is again a self-

contradiction. Thus, there exists a point  $x' \in \mathcal{S}_{x^0, \alpha_1}$  such that

$$\rho_1(x') = \rho(x^0). \quad \dots (3.16)$$



Now, taking (3.15) in particular for  $x = x^0$ , and realizing that  $N$  could be chosen suitably for any prescribed point  $x^0$ , we see that we have

$$\rho_1(x) \leq \rho(x), \quad x \in \Omega. \quad \dots (3.17)$$

Applying (3.17) in particular to the point  $x'$ , we have

$$\rho_1(x') \leq \rho(x'). \quad \dots (3.18)$$

Let  $\mathcal{S}_{x', \beta} \subseteq \mathcal{S}_{x^0, \alpha_1}$ . Then by (3.15),  $\max_{x \in \mathcal{S}_{x', \beta}} \rho_1(x) \leq \rho(x^0)$ , and therefore,

by Definition 3.2,

$$\rho(x') \leq \rho(x^0). \quad \dots (3.19)$$

Combining (3.16), (3.18) and (3.19) gives

$$\rho(x') = \rho_1(x') = \rho(x^0). \quad \dots (3.20)$$

Hence, in particular,  $x'$  is a regular point. And since  $x' \in \mathcal{S}_{x^0, \alpha_1} \subseteq N$ , we have hereby proved the everywhere denseness of  $R$  in  $\Omega$ .

To complete the proof of the lemma, let  $x'$  be any regular point. By the same argument that led to (3.15) above for the point  $x^0$ , we have in the present case that there is a sphere  $\mathcal{S}_{x', \alpha_1}$  such that

$$\rho_1(x) \leq \rho(x'), \quad x \in \mathcal{S}_{x', \alpha_1}. \quad \dots (3.21)$$

On the other hand, there is a minor of order  $\rho_1(x')$ , of some matrix (3.1), which is nonvanishing at  $x'$ ; since the partial derivatives constituting this minor are continuous, there is in fact a sphere  $\mathcal{S}_{x', \alpha_2}$  at every point of which this minor is nonvanishing, and therefore

$$\rho_1(x) \geq \rho_1(x'), \quad x \in \mathcal{S}_{x', \alpha_2}. \quad \dots (3.22)$$

If  $\mathcal{S}_{x', \alpha}$  denotes the smaller of the two spheres  $\mathcal{S}_{x', \alpha_1}$  and  $\mathcal{S}_{x', \alpha_2}$ , then, combining (3.21) and (3.22) with the fact that  $\rho(x') = \rho_1(x')$  (since  $x'$  is regular), we get

$$\rho_1(x) = \rho(x'), \quad x \in \mathcal{S}_{x', \alpha}. \quad \dots (3.23)$$

Thus, in particular,  $\rho_1$  is constant in  $\mathcal{S}_{x', \alpha}$ . It is easily verified with Definition 3.2 that in this circumstance the function  $\rho$  is likewise constant in  $\mathcal{S}_{x', \alpha}$ , and with the same value as  $\rho_1$ . Thus, we have

$$\rho(x) = \rho_1(x) = \rho(x'), \quad x \in \mathcal{S}_{x', \alpha}. \quad \dots (3.24)$$



# SUFFICIENT STATISTICS OF MINIMAL DIMENSION

This relationship shows immediately both (1) that all points of  $\mathcal{S}_{x',\alpha}$  are regular points—therefore  $R$  is an open set—and (2) that  $\lim_{x \rightarrow x'} \rho(x) = \rho(x')$ —therefore  $\rho$  is continuous on  $R$ .

This completes the proof of Lemma 3.1.

For completeness we include the following alternative characterization of the regular points of  $\Omega$ .

Lemma 3.2:  $R$  is precisely the set of points of continuity of the function  $\rho_1$ .

*Proof:* Since by the preceding lemma,  $\rho$  is continuous on the open set  $R$ , and  $\rho_1 \equiv \rho$  on  $R$  (by definition), it follows that every point of  $R$  is a continuity point of  $\rho_1$ .

Conversely, suppose  $x^0$  is a continuity point of  $\rho_1$ . Since  $\rho_1$  is integer-valued (thus, with discrete range), there is then a sphere  $\mathcal{S}_{x^0,\alpha}$  for every point  $x$  of which  $\rho_1(x) = \rho_1(x^0)$ . It therefore follows immediately from Definition 3.2 that  $\rho(x^0) = \rho_1(x^0)$ , that is,  $x^0 \in R$ . This completes the proof.

We are now ready to prove

Theorem 3.3: If  $x^0$  is a regular point of  $\Omega$ , there exists a  $\mathcal{P}$ -sufficient statistic,  $T$ , which is Euclidean of dimension  $\rho(x^0)$  at  $x^0$ , and regular at  $x^0$ .

*Proof:* If  $\rho(x^0) = n$ , the result follows immediately by taking  $T(x) \equiv x$ . We go on to consider the case  $\rho(x^0) < n$ .

For the regular point  $x^0$ , let us set, for brevity,  $r_0 = \rho(x^0) = \rho_1(x^0)$ . By Definition 3.1, there is a matrix (3.1) of which a particular  $r_0 \times r_0$  submatrix is nonsingular at  $x^0$ . Without loss of generality we may suppose that the  $r_0 \times r_0$  submatrix in question involves the first  $r_0$  coordinates of  $x$ . Thus, there exist points  $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(r_0)}$  in  $\Theta$ , and integers  $j_1, j_2, \dots, j_{r_0}$  each chosen from  $\{1, 2, \dots, v\}$ , such that the matrix

$$\Delta(x) = \left\| \left( \frac{\partial^2 \log p}{\partial x_i \partial \theta_{j_k}} \right)_{x, \theta^{(k)}} \right\|_{i, k = 1, 2, \dots, r_0} \quad \dots \quad (3.25)$$

is nonsingular at  $x^0$ . By virtue of the continuous differentiability of  $p$ , there is then a neighbourhood of  $x^0$ , at every point of which  $\Delta(x)$  is nonsingular; choose such a spherical neighbourhood, say  $\mathcal{S}_{x^0,\alpha}$ . But furthermore, since  $x^0$  is a regular point, it follows by Lemma 3.1 that we may choose  $\mathcal{S}_{x^0,\alpha}$  so small that

$$\rho(x) = \rho_1(x) = \rho_1(x^0), \quad x \in \mathcal{S}_{x^0,\alpha}; \quad \dots \quad (3.26)$$



that is, every point of  $\mathcal{S}_{x^0, \alpha}$  is a regular point and  $\rho$  is constant in  $\mathcal{S}_{x^0, \alpha}$ . So selecting  $\mathcal{S}_{x^0, \alpha}$ , we therefore have the following two facts : (1)  $\Delta(x)$  has nonvanishing determinant for all  $x \in \mathcal{S}_{x^0, \alpha}$ , and (2) for any point  $\theta \in \Theta$ , any integer  $j = 1, 2, \dots, \nu$  and any integer  $s = 1, 2, \dots, n$ , we have

$$\Delta(x) = \begin{vmatrix} \left( \frac{\partial^2 \log p}{\partial x_s \partial \theta_{j_1}} \right)_{x, \theta^{(1)}} & \vdots & \left( \frac{\partial^2 \log p}{\partial x_s \partial \theta_{j_{r_0}}} \right)_{x, \theta^{(r_0)}} \\ \vdots & \ddots & \vdots \\ \left( \frac{\partial^2 \log p}{\partial x_1 \partial \theta_j} \right)_{x, \theta} & \dots & \left( \frac{\partial^2 \log p}{\partial x_r \partial \theta_j} \right)_{x, \theta} & \left( \frac{\partial^2 \log p}{\partial x_s \partial \theta_j} \right)_{x, \theta} \end{vmatrix} = 0 \dots \quad (3.27)$$

Now, it follows from these facts, by a familiar argument with the Implicit Function Theorem, that there is a neighbourhood  $N \subseteq \mathcal{S}_{x^0, \alpha}$  of  $x^0$  such that, for each  $\theta \in \Theta$ , the functions

$$\left( \frac{\partial \log p}{\partial \theta_j} \right)_\theta, \quad j = 1, 2, \dots, \nu \quad \dots \quad (3.28)$$

are in  $N$ , continuously differentiable functions of only the functions  $\eta_k$ , defined by

$$\eta_k(x) = \left( \frac{\partial \log p}{\partial \theta_{j_k}} \right)_{x, \theta^{(k)}}, \quad k = 1, 2, \dots, r_0, \quad \dots \quad (3.29)$$

the neighbourhood  $N$  being such that its image under  $\eta = (\eta_1, \eta_2, \dots, \eta_{r_0})$  is an open sphere  $S$  in an  $r_0$ -dimensional Euclidean space  $E^{r_0}$ . In other words, there are functions  $F'_j$  on  $S \times \Theta$ , which are continuously differentiable in the coordinates  $y_1, y_2, \dots, y_{r_0}$  of a point  $y \in S$ , such that

$$\frac{\partial \log p}{\partial \theta_j} = F'_j(\eta(x), \theta), \quad x \in N, \quad \theta \in \Theta, \quad j = 1, 2, \dots, \nu. \quad \dots \quad (3.30)$$

It is a consequence of the relations (3.30) that there exist continuous functions  $F$ , on  $S \times \Theta$ , and  $G$ , on  $N$ , such that

$$\log p(x, \theta) = F(\eta(x), \theta) + G(x), \quad (x, \theta) \in N \times \Theta \quad \dots \quad (3.31)$$



and such that the derivatives

$$\left\{ \begin{array}{l} \frac{\partial F}{\partial y_k}, \quad k = 1, 2, \dots, r_0, \\ \frac{\partial F}{\partial \theta_j}, \quad j = 1, 2, \dots, v, \\ \frac{\partial^2 F}{\partial y_k \partial \theta_j}, \frac{\partial^2 F}{\partial \theta_j \partial y_k}, \quad k = 1, 2, \dots, r_0; j = 1, 2, \dots, v, \\ \frac{\partial G}{\partial x_i}, \quad i = 1, 2, \dots, n \end{array} \right. \quad \dots (3.32)$$

are all continuous in their respective domains  $S \times \Theta$  and  $N$ . A sketch will suffice to show how this result may be established. From (3.30) for  $j = 1$ , it follows that

$$\log p(x, \theta) = F_1(\eta(x), \theta) + G_1(x, \theta_2, \theta_3, \dots, \theta_v), \quad \dots (3.33)$$

where  $F_1$  is an integral, with respect to  $\theta_1$ , of  $F'_1$ , and so has the continuous differentiability properties described for  $F$  in (3.31) and (3.32). Since  $F_1$  has these properties, and  $\eta$  is continuously differentiable, and since  $\log p$  also has the continuous differentiability properties, it is implied by (3.33) that  $G_1$  likewise has these properties. Differentiating (3.33) with respect to  $\theta_2$  and equating the result to the right-hand side of (3.30) for  $j = 2$ , we obtain

$$\frac{\partial G_1}{\partial \theta_2} = F'_2 - \frac{\partial F_1}{\partial \theta_2}. \quad \dots (3.34)$$

The right-hand side of this equation depends on  $x$  only through  $\eta$ , and is, moreover, independent of  $\theta_1$  since the left-hand side is. Therefore, (3.34) asserts that

$$G_1(x, \theta_2, \theta_3, \dots, \theta_v) = \hat{F}_2(\eta(x), \theta_2, \theta_3, \dots, \theta_v) + G_2(x, \theta_3, \theta_4, \dots, \theta_v), \quad \dots (3.35)$$

where  $\hat{F}_2$  and  $G_2$  again have the stated properties of continuous differentiability. Inserting the right-hand side of (3.35) into (3.33), and setting  $F_2 = F_1 + \hat{F}_2$ , we obtain

$$\log p(x, \theta) = F_2(\eta(x), \theta) + G_2(x, \theta_3, \theta_4, \dots, \theta_v). \quad \dots (3.36)$$

Continuing in the manner indicated through all the conditions (3.30), we ultimately obtain (3.31) and (3.32) as asserted.

Define the function  $T$  on  $\Omega$  as follows:

$$T(x) = \begin{cases} \eta(x); & x \in N, \\ x, & x \in \Omega - N. \end{cases} \quad \dots (3.37)$$



$\mathcal{R}_T$ , the range of  $T$ , is  $S \cup (\Omega - N)$ . On  $\mathcal{R}_T \times \Theta$  define the function  $f$  as follows :

$$f(u, \theta) = \begin{cases} e^{F(u, \theta)}, & (u, \theta) \in S \times \Theta \\ p(u, \theta), & (u, \theta) \in (\Omega - N) \times \Theta \end{cases} \quad \dots (3.38)$$

and define the function  $g$  on  $\Omega$  by

$$g(x) = \begin{cases} e^{G(x)}, & x \in N, \\ 1, & x \in \Omega - N. \end{cases} \quad \dots (3.39)$$

Then, by virtue of (3.31), we have

$$p(x, \theta) = f(T(x), \theta) g(x) \quad (x, \theta) \in \Omega \times \Theta. \quad \dots (3.40)$$

This relation will verify the sufficiency of  $T$  as soon as we demonstrate the requisite measurability properties of  $f$  and  $g$ .

Regarding  $f$ , we must show, for each fixed  $\theta$ , that the inverse image, under  $f(T(\cdot), \theta)$ , of a Borel set on the real line is a Lebesgue set in  $\Omega$ . Let  $H$  be a Borel set on the real line, and define the sets (for a particular, fixed  $\theta$ )

$$\begin{aligned} A_1 &= \{u \in S \mid e^{F(u, \theta)} \in H\}, \\ A_2 &= \{u \in \Omega - N \mid p(u, \theta) \in H\}. \end{aligned} \quad \dots (3.41)$$

Then,

$$\begin{aligned} \text{inverse image of } H \text{ under } f(T(\cdot), \theta) &= T^{-1}(A_1 \cup A_2) \\ &= \eta^{-1}(A_1) \cup A_2. \end{aligned} \quad \dots (3.42)$$

Since  $e^{F(\cdot, \theta)}$  and  $p(\cdot, \theta)$  are continuous functions on  $S$  and  $\Omega - N$ , respectively, the sets  $A_1$  and  $A_2$  are Borel sets. Also,  $\eta$  is a continuous function, and therefore  $\eta^{-1}(A_1)$  is a Borel set. Hence, the union of  $\eta^{-1}(A_1)$  and  $A_2$  is a Borel set in  $\Omega$ , and we have, by virtue of (3.42), proved the Lebesgue (in fact, the Borel) measurability of  $f(T(\cdot), \theta)$ .

From the Lebesgue measurability of  $p(\cdot, \theta)$  and  $f(T(\cdot), \theta)$  in (3.40), it follows that  $g$  is Lebesgue measurable. Hence, Lemma 1.1 asserts that  $T$  is a sufficient statistic. In the neighbourhood  $N$  of  $x^0$ ,  $T = \eta$ ; and  $\eta$  is continuously differentiable on  $N$  into  $E^{r_0}$ , and its Jacobian matrix is of rank  $r_0$  at  $x^0$  (the matrix  $\Delta(x)$  (see (3.25)), which is nonsingular at  $x^0$ , is an  $r_0 \times r_0$  submatrix of the Jacobian matrix of  $\eta$ ). Thus,  $T$  is Euclidean of dimension  $\rho(x^0)$  at  $x^0$ , and regular at  $x^0$ .

Theorem 3.3 is therefore fully proved.

We remark that (3.29) constitutes an *explicit* characterization of the function  $\eta$  with which we have constructed the sufficient statistic  $T$  in (3.37). That is, our proof is explicitly constructive.



# SUFFICIENT STATISTICS OF MINIMAL DIMENSION

## 4. GLOBAL ACHIEVEMENT OF MINIMAL DIMENSIONALITY

Theorem 3.3 gives rise to the question: for how "big" a subset of  $R$  is it possible to achieve, in a single sufficient statistic, the local minimal dimension,  $\rho(x)$ , at all points  $x$  of the subset? The next theorem provides an answer.

**Theorem 4.1:** *There exists a  $\mathcal{P}$ -sufficient statistic,  $T^*$ , which, for almost all (Lebesgue) regular points  $x^0$  of  $\Omega$ , is Euclidean of dimension  $\rho(x^0)$  at  $x^0$ , and regular at  $x^0$ .*

*Proof:* The key to this result lies in the observation of what has actually been established in the proof of the preceding theorem. It will be seen, on re-examining the details, that, in fact, the sufficient statistic  $T$  which we constructed has the following property: every point  $x$  of the neighbourhood  $N$  is a regular point, and for each  $x \in N$ ,  $T$  is Euclidean of dimension  $\rho(x)$  at  $x$ , and regular at  $x$ . We shall employ the full force of this to prove the present theorem.

The set  $R$  is an open set, and it is therefore a denumerable union of mutually disjoint cells :

$$R = \bigcup_{s=1}^{\infty} I_s. \quad \dots \quad (4.1)$$

(A cell is a bounded interval with the upper faces included and the lower faces excluded.) Moreover, we may take such a representation (4.1) that, for each cell  $I_s$ , its closure,  $\bar{I}_s$ , is a subset of  $R$ .

Now our observation above, concerning the full implication of the proof of Theorem 3.3, enables us here to assert the following, for each  $s = 1, 2, \dots$  : for each point  $x^0 \in \bar{I}_s$  there is an open interval  $K_{s, x^0}$ , centered at  $x^0$ , lying within  $R$ , and a sufficient statistic  $T_{s, x^0}$  which is Euclidean of dimension  $\rho(x)$  at each point  $x \in K_{s, x^0}$ , and regular at each point  $x \in K_{s, x^0}$ . By the Heine-Borel Theorem, there is a finite subcollection of the collection  $\{K_{s, x^0}, x^0 \in \bar{I}_s\}$  which covers  $\bar{I}_s$ . Let this finite, covering subcollection be  $\{K'_{s1}, K'_{s2}, \dots, K'_{st_s}\}$ , and the corresponding  $T_{t, x^0}$ 's be  $T'_{s1}, T'_{s2}, \dots, T'_{st_s}$ . Then, by a well-known procedure, using the bounding hyperplanes of the  $K'_{si}$  in which to define new faces where necessary, we are able to designate a finite collection of mutually disjoint cells,  $\{K_{s1}, K_{s2}, \dots, K_{st_s}\}$ , each of which is a subset of some  $K'_{si}$ , and such that

$$\bigcup_{i=1}^{t_s} K_{si} = I_s. \quad \dots \quad (4.2)$$

For each  $i = 1, 2, \dots, t_s$ , let  $T_{si}$  be some particular one of those  $T'_{sj}$  such that  $K_{si} \subseteq K'_{sj}$ . Thus, for each  $i = 1, 2, \dots, t_s$ , the sufficient statistic  $T_{si}$  is Euclidean of dimension  $\rho(x)$  at each point  $x \in K_{si} \equiv \text{interior of } K_{si}$ , and is regular at each  $x \in K_{si}^0$ .



Consider the above carried out for all  $s = 1, 2, \dots$ . Let the functions  $f_{si}$  and  $g_{si}$  (with the continuous differentiability properties found in the proof of Theorem 3.3) be such that

$$p(x, \theta) = f_{si}(T_{si}(x), \theta) g_{si}(x), \quad (x, \theta) \in K_{si} \times \theta; \quad \dots \quad (4.3)$$

$$i = 1, 2, \dots, t_s,$$

$$s = 1, 2, \dots$$

In particular, for each pair  $s, i$  such that the (constant) value of  $\rho$  on  $K_{si}$  is  $n$ , we take  $T_{si}(x) \equiv x$ ,  $f_{si}(x, \theta) = p(x, \theta)$  and  $g_{si}(x) \equiv 1$ . (This is in agreement with our definition of  $T$  for  $\rho(x^0) = n$  in the proof of Theorem 3.3).

For each of the integers  $r = 1, 2, \dots, n-1$  in the range of the function  $\rho$  on  $R$ , let  $E^r$  be a fixed  $r$ -dimensional Euclidean space. For every index pair  $s, i$  such that  $T_{si}$  is Euclidean of dimension  $r$ , we shall consider  $D_{si}$ , the range of  $K_{si}$  under  $T_{si}$ , as a set in  $E^r$ . Since our present statistics  $T_{si}$  are of the kind constructed in the proof of Theorem 3.3 we have that the sets  $D_{si}$  are all bounded. It follows, then, by virtue of this boundedness and the denumerability of the collection of the  $D_{si}$ , that there exist translation vectors  $v_{si}$ , in the respective imbedding spaces  $E^r$ ,  $r = 1, 2, \dots, n-1$ , of the  $D_{si}$  of positive dimension  $< n$  such that the following holds: if  $D_{si}^*$  denotes the translation of  $D_{si}$  by  $v_{si}$ , then for any two sets  $D_{si}$  and  $D_{s'i'}$  in the same  $E^r$ , the sets  $D_{si}^*$  and  $D_{s'i'}^*$  are disjoint.

The cases in which  $T_{si}$  is Euclidean of dimension 0 are those in which the  $T_{si}$  are constant. The  $D_{si}$  for such a  $T_{si}$  is (a set consisting of) a single number. For these cases we can likewise define numbers  $v_{si}$  such that no two of the numbers  $D_{si}^* = D_{si} + v_{si}$  are equal.

With the  $v_{si}$  defined as here described for the  $T_{si}$  of dimensions 0, 1, 2, ...,  $n-1$ , and with  $v_{si} = \text{null vector in } E_0^n$  for a  $T_{si}$  of dimension  $n$ , we define

$$T_{si}^*(x) = T_{si}(x) + v_{si}, \quad x \in K_{si}; \quad \dots \quad (4.4)$$

$$i = 1, 2, \dots, t_s,$$

$$s = 1, 2, \dots$$

and

$$f_{si}^*(u, \theta) = f_{si}(u - v_{si}, \theta), \quad (u, \theta) \in D_{si}^* \times \theta; \quad \dots \quad (4.5)$$

$$i = 1, 2, \dots, t_s,$$

$$s = 1, 2, \dots$$



# SUFFICIENT STATISTICS OF MINIMAL DIMENSION

We are now finally able to make the crucial definitions :

$$T^*(x) = \begin{cases} T_{si}^*(x), & x \in K_{si}; \quad i = 1, 2, \dots, t_s; \quad s = 1, 2, \dots, \\ x, & x \in \Omega - R, \end{cases} \quad \dots \quad (4.6)$$

$$f^*(u, \theta) = \begin{cases} f_i^*(u, \theta), & (u, \theta) \in D_{si}^* \times \Theta; \quad i = 1, 2, \dots, t_s; \quad s = 1, 2, \dots, \\ p(u, \theta), & (u, \theta) \in (\Omega - R) \times \Theta, \end{cases} \quad \dots \quad (4.7)$$

and 
$$g^*(x) = \begin{cases} g_{si}(x), & x \in K_{si}; \quad i = 1, 2, \dots, t_s; \quad s = 1, 2, \dots, \\ 1, & x \in \Omega - R. \end{cases} \quad \dots \quad (4.8)$$

The first of these definitions specifies a function  $T^*$  over all of  $\Omega$  since  $R = \bigcup_{s,i} K_{si}$ . The range of  $T^*$ , say  $\mathcal{R}_{T^*}$ , is clearly  $(\bigcup_{s,i} D_{si}^*) \cup (\Omega - R)$ , so that  $f^*$  is a function defined on  $\mathcal{R}_{T^*} \times \Theta$ . It is for purposes of definition of this function  $f^*$  that it has been necessary to introduce the disjoint translations of the sets  $D_{si}$ . We could not directly define, for all  $i$  and  $s$ ,  $f^*(u, \theta) = f_{si}(u, \theta)$  for  $u \in D_{si}$ ,  $\theta \in \Theta$ , because if two sets  $D_{si}$  and  $D_{s'i'}$  have a nonempty intersection, there is no assurance that the functions  $f_{si}$  and  $f_{s'i'}$  agree on this intersection. We may remark that it was not necessary to translate those  $D_{si}$  for which  $T_{si}$  is of dimension  $n$ , since in these cases we have taken  $T_{si} = \text{identity function on } \Omega$  and therefore  $D_{si} = K_{si}$ , and the  $K_{si}$  are already mutually disjoint and disjoint from  $\Omega - R$ .

It is an immediate consequence of (4.3)–(4.8) that

$$p(x, \theta) = f^*(T^*(x), \theta) \cdot g^*(x), \quad (x, \theta) \in \Omega \times \Theta. \quad \dots \quad (4.9)$$

We need not carry out explicitly the demonstration of the Lebesgue (in fact, Borel) measurability of the functions  $f^*(T^*(\cdot), \theta)$  and  $g^*$ . The proof runs along the same lines as the proof of measurability of the function  $f(T(\cdot), \theta)$  in (3.40). In the present case there will be, in place of the single Borel set  $\eta^{-1}(A_1)$  in (3.42), a denumerable union of such Borel sets. And one will have to employ here the fact that translations are Borel measurable functions.

Hence, (4.9) fulfills the conditions of Lemma 1.1, and  $T^*$  is a sufficient statistic. Moreover, for each  $x \in \bigcup_{s,i} K_{si}^0$ ,  $T^*$  is Euclidean of dimension  $\rho(x)$  at  $x$ , and regular at  $x$  (obviously, this property of the  $T_{si}$  in  $K_{si}^0$  is preserved under the translations (4.4)). It remains, then, only to observe that the complement in  $R = \bigcup_{s,i} K_{si}$  of the set  $\bigcup_{s,i} K_{si}^0$  is a union of subsets of denumerably many hyperplanes (that is, a union of faces of denumerably many cells), and so is of  $n$ -dimensional Lebesgue measure 0.

This completes the proof of Theorem 4.1.



The theorem we have just proved concerns global achievement of the *local* minimal dimensions of a sufficient statistic. It does not deal with a globally defined minimal dimension. On the other hand, it is customary practice in the literature to refer to a global minimal dimension. Specifically, reference is to "the smallest number of continuously differentiable, real-valued functions on  $\Omega$  which can constitute a sufficient statistic." This particular definition of global minimal dimension is certainly more strict than it need be, in requiring continuous differentiability throughout  $\Omega$ . A more useful definition, in the light of our results, would be the following one :

*Definition 4.1:* The (almost sure) global minimal dimension of a  $\mathcal{P}$ -sufficient statistic—to be denoted by  $d_0$ —is the smallest integer  $r$  such that the following is true: there exists a  $\mathcal{P}$ -sufficient statistic  $T$  and an open set  $A \subseteq \Omega$ , with  $\Omega - A$  of Lebesgue measure 0, such that  $T$  is Euclidean of dimension  $r$  at every point of  $A$ , and is continuously differentiable throughout  $A$ .

The following theorem now expresses the implications of our above results regarding the number  $d_0$  and the construction of a sufficient statistic of global minimal dimension.

*Theorem 4.2:* The global minimal dimension  $d_0$  satisfies the inequality

$$d_0 \geq \max_{x \in R} \rho(x). \quad \dots (4.10)$$

There exists a  $\mathcal{P}$ -sufficient statistic,  $T^+$ , which is Euclidean of dimension  $\max_{x \in R} \rho(x)$  at every point of an open set  $A \subseteq R$ , with  $R - A$  of Lebesgue measure 0, and which is continuously differentiable throughout  $A$ .

Hence, if  $\Omega - R$  is of Lebesgue measure 0, then equality holds in (4.10), and  $T^+$  is a  $\mathcal{P}$ -sufficient statistic of global minimal dimension.

*Proof:* The inequality (4.10) is an immediate consequence of Theorem 3.2.

A statistic  $T^+$  of the kind asserted in the second statement in the theorem can be obtained by a simple modification of the statistic  $T^*$  constructed in the proof of Theorem 4.1. And, like  $T^*$ , this new statistic will be Euclidean and continuously differentiable on  $\bigcup_{s,i} K_{si}^0$ , which is a subset of  $R$  whose complement in  $R$  is of Lebesgue measure 0.

Since the idea underlying the construction of  $T^+$  out of  $T^*$  is quite simple, whereas the details are cumbersome, we shall content ourselves with merely indicating how it is to be done. For brevity, set  $r_1 = \max_{x \in R} \rho(x)$ . If  $T_{si}^*$  is of dimension  $r < r_1$  over  $K_{si}$ , write in more detail,

$$T_{si}^*(x) = (h_{si1}(x), h_{si2}(x), \dots, h_{sir}(x)), \quad \dots (4.11)$$

where the  $h_{sij}$  are the continuously differentiable component functions of  $T_{si}^*$ . Let  $c_{si1}, c_{si2}, \dots, c_{si, r_1-1}$  be  $r_1 - r$  constants, and define

$$T_{si}^{*+}(x) = (h_{si1}(x), h_{si2}(x), \dots, h_{sir}(x), c_{si1}, c_{si2}, \dots, c_{si, r_1-r}), \quad x \in K_{si}. \quad \dots (4.12)$$



# SUFFICIENT STATISTICS OF MINIMAL DIMENSION

Thus, by this device, the range of each  $T_{si}^{*+}$  on its  $K_{si}$  is a (bounded) set in a Euclidean  $r_1$ -space; we view all these ranges as lying in a common  $r_1$ -dimensional space. By following this up with the device of translations, as in the proof of the preceding theorem—and, in the case  $r_1 = n$ , possibly also contraction transformations (minifications relative to suitable points) if they are needed—we are able to arrive at functions  $T_{si}^+$  on the  $K_{si}$ , which have mutually disjoint ranges, and ranges which are, specifically, in the case  $r_1 = n$ , disjoint from  $\Omega - R$ . We then define

$$T^+(x) = \begin{cases} T_{si}^+(x), & x \in K_{si}, i = 1, 2, \dots, t_s; s = 1, 2, \dots, \\ x, & x \in \Omega - R, \end{cases} \quad \dots \quad (4.13)$$

and we are enabled also, by the disjointness of the ranges of the  $T_{si}^+$  on the  $K_{si}$ , to define, as in the preceding theorem, functions  $f^+$  and  $g^+$  which verify that  $T^+$  is a sufficient statistic. A  $T_{si}^+$ , being formed from  $T_{si}^{*+}$  by a translation and (possibly) a contraction, is of the form  $aT_{si}^{*+} + b$  for some constants  $a$  and  $b$ . Hence, it is clear that  $T_{si}^+$  is continuously differentiable in  $K_{si}^0$ . And therefore, since, also, each  $T_{si}^+$  is Euclidean of dimension  $r_1$  in  $K_{si}$ , we have, by (4.13), that  $T^+$  is Euclidean of dimension  $r_1$  at each point of  $\bigcup_{s,i} K_{si}^0$ , and continuously differentiable throughout this set.

The last assertion in the statement of Theorem 4.2 is obvious, and the proof of the theorem is therefore complete.

Finally, we say a word about the relation between *dimensional* minimality of  $\mathcal{P}$ -sufficient statistics and *functional* minimality of such statistics. Lehmann and Scheffé (1950) examined the notion which they termed "minimality" of a sufficient statistic; we shall here designate this as "functional minimality." A sufficient statistic for our family  $\mathcal{P}$  is a functionally minimal one if it is almost everywhere (Lebesgue) in  $\Omega$  a function of any other sufficient statistic. The authors mentioned have shown, in their Theorem 6.3, that for our present family  $\mathcal{P}$  there does exist a functionally minimal sufficient statistic. The question therefore arises: how far toward functional minimality can we arrive by choosing a dimensionally minimal (local or global) sufficient statistic? An answer that we are able to give fairly readily is expressed by the following theorem.

**Theorem 4.3:** *The  $\mathcal{P}$ -sufficient statistics  $T^*$  and  $T^+$ , of the preceding two theorems, are locally functionally minimal at almost all (Lebesgue) points of  $R$ . That is, for each point  $x$  in a set  $A \subseteq R$ , with  $R - A$  of measure 0, there is a neighbourhood  $N$  of  $x$  such that if  $T$  is any  $\mathcal{P}$ -sufficient statistic, then  $T^*$  and  $T^+$  are functions of  $T$  in  $N$ .*

*Proof:* The set  $A$  in question may, again, be taken to be  $\bigcup_{s,i} K_{si}^0$ . In fact, what is true is that (referring to the elements in the proof of Theorem 4.1) for each pair  $s, i$ , the statistic  $T_{si}$  is, on  $K_{si}$ , a function of any other sufficient statistic. Once this is established, the asserted local functional minimality of  $T^*$  then follows from (4.4) and (4.6).



To establish our assertion concerning  $T_{si}$ , we note that in  $K_{si}$  this statistic has, for its component functions, a set of functions of the form (3.29). Thus, we have to show that each of the functions

$$\left( \frac{\partial \log p}{\partial \theta_{jk}} \right)_{x, \theta^{(k)}} \quad \dots \quad (4.14)$$

in question is a function of any sufficient statistic. Let  $T$  be any particular sufficient statistic. Then, by Lemma 1.1, we have the representation (1.6) for some functions  $f$  and  $g$ . It follows that

$$\left( \frac{\partial \log p}{\partial \theta_{jk}} \right)_{x, \theta^{(k)}} = \left( \frac{\partial \log f(T(x), \cdot)}{\partial \theta_{jk}} \right)_{\theta^{(k)}}, \quad \dots \quad (4.15)$$

and this shows immediately that the functions (4.14) are functions of  $T$ .

This completes the proof of the assertion of Theorem 4.3 for  $T^*$ .

To prove the assertion for  $T^+$ , we have merely to note that on  $K_{si}$  each component of  $T_{si}^+$  is either a linear function of the corresponding component of  $T_{si}^*$ , or a constant. From this, by virtue of the result already established for  $T^*$ , it follows that  $T^+$  is a function of any particular sufficient statistic, in each  $K_{si}$ . This proves the asserted result for  $T^+$ .

We have therefore completed the proof of Theorem 4.3.

The question remains open as to when we can assert the existence of a sufficient statistic which is continuously differentiable in an open set  $A \subseteq R$ , with  $R-A$ , of Lebesgue measure 0, and which is functionally minimal in  $A$ .

## 5. TWO ILLUSTRATIVE EXAMPLES

In this section we shall apply the preceding results in two examples. The first example pertains to the well-known Behrens-Fisher problem; here we show that  $\Omega - R$  has Lebesgue measure 0 and that  $\rho(x) \equiv 4$  for  $x \in R$ . The second example exhibits a problem where again  $\Omega - R$  has Lebesgue measure 0 but  $\rho(x)$  is not constant for  $x \in R$ .

Consider  $m+n$ , ( $m, n \geq 2$ ), independent random variables  $X_1, \dots, X_m, Y_1, \dots, Y_n$ , where  $X_i$  is a normal random variable with mean  $\theta_1$  and variance  $\theta_2^2$  for  $i = 1, \dots, m$  and  $Y_j$  is normal with mean  $\theta_3$  and variance  $\theta_4^2$  for  $j = 1, \dots, n$ . Let  $\mathcal{P} = \{\mu_\theta : \theta \in \Theta\}$  be the family of measures generated by the joint density of  $X_1, \dots, Y_n$ , i.e.,

$$\left\{ \begin{array}{l} \theta = (\theta_1, \theta_2, \theta_3, \theta_4) \\ \Theta = (-\infty, \infty) \times (0, \infty) \times (-\infty, \infty) \times (0, \infty) \\ \Omega = E^{m+n} \\ \mu_\theta(A) = \int_A p(x_1, \dots, y_n; \theta) dx_1 \dots dy_n, \quad A \text{ a Lebesgue set} \end{array} \right. \quad \dots \quad (5.1)$$



# SUFFICIENT STATISTICS OF MINIMAL DIMENSION

where  $p(x_1, \dots, y_n; \theta)$  denotes the joint density of  $X_1, \dots, Y_n$ . It is clear that conditions (1.1) and (1.2) are met and we proceed immediately to the calculation of  $\rho_1(x)$  and  $\rho(x)$ . The calculation of the various mixed partial derivatives is routine but we list the results for the sake of completeness.

$$\left\{ \begin{array}{l} \frac{\partial^2 \log p}{\partial x_i \partial \theta_3} = \frac{\partial^2 \log p}{\partial y_j \partial \theta_1} = \frac{\partial^2 \log p}{\partial x_i \partial \theta_4} = \frac{\partial^2 \log p}{\partial y_j \partial \theta_2} = 0 \quad \begin{array}{l} i = 1, \dots, m \\ j = 1, \dots, n \end{array} \\ \\ \frac{\partial^2 \log p}{\partial x_i \partial \theta_2} = \frac{2}{\theta_2^3} (x_i - \theta_1) \quad i = 1, \dots, m \\ \\ \frac{\partial^2 \log p}{\partial x_i \partial \theta_1} = \frac{1}{\theta_2^2} \quad i = 1, \dots, m \\ \\ \frac{\partial^2 \log p}{\partial y_j \partial \theta_3} = \frac{1}{\theta_4^2} \quad j = 1, \dots, n \\ \\ \frac{\partial^2 \log p}{\partial y_j \partial \theta_4} = \frac{2}{\theta_4^3} (y_j - \theta_3) \quad j = 1, \dots, n. \end{array} \right. \quad \dots (5.2)$$

Thus it is clear that

$$\text{rank } L(x; j_1, \dots, j_{m+n}; \theta^{(1)}, \dots, \theta^{(m+n)}) \leq 4. \quad \dots (5.3)$$

Let  $j_1 = 1, j_2 = 2, j_3 = 3, j_4 = 4$  and  $j_5 = \dots = j_{m+n} = 1$ , then

$$L(x; 1, 2, 3, 4, 1, \dots, 1; \theta^{(1)}, \dots, \theta^{(m+n)}) =$$

$$\left\| \begin{array}{cccccc} \frac{1}{\theta_2^{(1)^2}} & \frac{2}{\theta_2^{(2)^3}} & (x_1 - \theta_1^{(2)}) & 0 & 0 & \frac{1}{\theta_2^{(5)^2}} \dots \frac{1}{\theta_2^{(m+n)^2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\theta_2^{(1)^2}} & \frac{2}{\theta_2^{(2)^3}} & (x_m - \theta_1^{(2)}) & 0 & 0 & \frac{1}{\theta_2^{(5)^2}} \dots \frac{1}{\theta_2^{(m+n)^2}} \\ 0 & 0 & \frac{1}{\theta_4^{(3)^2}} & \frac{2}{\theta_4^{(4)^3}} & (y_1 - \theta_3^{(4)}) & 0 \dots 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \frac{1}{\theta_4^{(3)^2}} & \frac{2}{\theta_4^{(4)^3}} & (y_n - \theta_3^{(4)}) & 0 \dots 0 \end{array} \right\| \quad \dots (5.4)$$



Hence, defining the set  $N$  by

$$N = \{x | x_1 = x_2 = \dots = x_m \text{ or } y_1 = y_2 = \dots = y_n\}, \quad \dots (5.5)$$

we see that for  $x \in \Omega - N$  we have

$$\text{rank } L(x; 1, 2, 3, 4, 1, \dots, 1 : \theta^{(1)}, \dots, \theta^{(m+n)}) = 4 \quad \dots (5.6)$$

$$\text{and} \quad \rho(x) = \rho_1(x), \quad \dots (5.7)$$

$$\text{while for } x \in N \quad \rho(x) > \rho_1(x). \quad \dots (5.8)$$

$$\text{Next we note that} \quad \mu_\theta(N) \equiv 0 \quad \dots (5.9)$$

and since  $\mu_\theta$  is absolutely continuous with respect to Lebesgue measure and conversely, it follows that the Lebesgue measure of  $N$  is also 0. From (5.7) and (5.8) we have that

$$\Omega - R = N \quad \dots (5.10)$$

and hence we have shown that  $R$  is an everywhere dense set whose complement has Lebesgue measure 0. Finally we define the function

$$T(x) = \left\{ \sum_{i=1}^m x_i, \sum_{i=1}^m x_i^2, \sum_{j=1}^n y_j, \sum_{j=1}^n y_j^2 \right\} \text{ for } x \in \Omega \quad \dots (5.11)$$

and note that with  $j_1 = 1, j_2 = 2, j_3 = 3, j_4 = 4$  and  $\theta^{(1)} = \theta^{(2)} = \theta^{(3)} = \theta^{(4)} = (0, 1, 0, 1)$  for  $x \in R$  this is, up to additive constants, exactly the function defined in the proof of Theorem 3.3; thus  $T(x)$  is a sufficient statistic for the family  $\mathcal{P}$  defined by (5.1) which is regular everywhere on  $R$  and of minimum dimension everywhere on  $R$ .

For the second example we consider two independent random variables  $X_1$  and  $X_2$  with density functions  $q(\xi, \theta_1)$  and  $q(\xi, \theta_2)$  where  $\theta_i \in (0, \infty)$ , ( $i = 1, 2$ ) and  $q(\xi, \theta_i)$  is defined as follows :

$$q(\xi, \theta_i) = \begin{cases} \frac{1}{1 + \theta_i \sqrt{2\pi}} e^{-\xi^2/2\theta_i^2} & \xi < 0 \\ \frac{1}{1 + \theta_i \sqrt{2\pi}} & 0 \leq \xi \leq 1 \\ \frac{1}{1 + \theta_i \sqrt{2\pi}} e^{-(\xi-1)^2/2\theta_i^2} & \xi > 1, \end{cases} \quad \dots (5.12)$$

$i = 1, 2$ . The family of measures  $\mathcal{P} = \{\mu_\theta : \theta \in \Theta\}$  considered is that generated by the



# SUFFICIENT STATISTICS OF MINIMAL DIMENSION

joint density of  $X_1$  and  $X_2$ , i.e.,

$$\left\{ \begin{array}{l} \theta = (\theta_1, \theta_2) \\ \Theta = (0, \infty) \times (0, \infty) \\ \Omega = E^2 \\ \mu_\theta(A) = \int_A p(x_1, x_2; \theta) dx_1 dx_2, \quad A \text{ a Lebesgue set} \end{array} \right. \quad \dots \quad (5.13)$$

where  $p(x_1, x_2; \theta)$  is equal to  $q(x_1, \theta_1) q(x_2, \theta_2)$ . Clearly  $p(x; \theta)$  is positive and continuous at all points of  $\Omega \times \Theta$  and an easy calculation verifies the existence and continuity of all the necessary partial derivatives.

Now we define the sets

$$\left\{ \begin{array}{l} A_1 = \{x | x_1, x_2 < 0\} \\ A_2 = \{x | x_1 < 0, x_2 > 1\} \\ A_3 = \{x | x_1 > 1, x_2 < 0\} \\ A_4 = \{x | x_1, x_2 > 1\} \\ A_5 = \{x | x_1 < 0, 0 \leq x_2 \leq 1\} \\ A_6 = \{x | 0 \leq x_1 \leq 1, x_2 < 0\} \\ A_7 = \{x | 0 \leq x_1 \leq 1, x_2 > 1\} \\ A_8 = \{x | x_1 > 1, 0 \leq x_2 \leq 1\} \\ A_9 = \{x | 0 \leq x_1, x_2 \leq 1\}. \end{array} \right. \quad \dots \quad (5.14)$$

Clearly  $\Omega = \bigcup_{i=1}^9 A_i$ , and minimum dimension will be computed by considering the problem in individual  $A_i$ 's. Let  $x \in A_1$ ; then

$$p(x; \theta) = \frac{1}{(1+\theta_1\sqrt{2\pi})(1+\theta_2\sqrt{2\pi})} e^{-x_1^2/2\theta_1^2 - x_2^2/2\theta_2^2} \quad \dots \quad (5.15)$$

and

$$\left\{ \begin{array}{l} \frac{\partial^2 \log p}{\partial x_i \partial \theta_i} = \frac{2}{\theta_i^3} \quad i = 1, 2, \\ \frac{\partial^2 \log p}{\partial x_i \partial \theta_j} = 0 \quad i \neq j; \quad i, j = 1, 2. \end{array} \right. \quad \dots \quad (5.16)$$



Thus

$$\text{rank } L(x; j_1 j_2; \theta^{(1)}, \theta^{(2)}) \leq 2 \text{ for } x \in A_1. \quad \dots (5.17)$$

Choose  $j_1 = 1$  and  $j_2 = 2$ , then

$$L(x; 1, 2; \theta^{(1)}, \theta^{(2)}) = \begin{vmatrix} \frac{2x_1}{\theta_1^{(1)^3}} & 0 \\ 0 & \frac{2x_2}{\theta_2^{(2)^3}} \end{vmatrix} \quad \dots (5.18)$$

and hence

$$\rho(x) = \rho_1(x) = 2, \quad x \in A_1. \quad \dots (5.19)$$

Similar reasoning shows that

$$\rho(x) = \rho_1(x) = 2 \quad x \in A_2 \cup A_3 \cup A_4 \quad \dots (5.20)$$

and we therefore obtain

$$\bigcup_{i=1}^4 A_i \subset R. \quad \dots (5.21)$$

Now consider  $x \in$  interior  $A_5$ , then

$$p(x; \theta) = \frac{1}{(1 + \theta_1 \sqrt{2\pi})(1 + \theta_2 \sqrt{2\pi})} e^{-x_1^2/2\theta_1^2} \quad \dots (5.22)$$

and

$$\begin{cases} \frac{\partial^2 \log p}{\partial x_1 \partial \theta_1} = \frac{2x_1}{\theta_1^3} \\ \frac{\partial^2 \log p}{\partial x_2 \partial \theta_2} = \frac{\partial^2 \log p}{\partial x_i \partial \theta_j} = 0 \quad i \neq j; \quad i, j = 1, 2. \end{cases} \quad \dots (5.23)$$

In this case

$$\text{rank } L(x; j_1, j_2; \theta^{(1)}, \theta^{(2)}) \leq 1 \quad \dots (5.24)$$

and

$$L(x; 1, 2; \theta^{(1)}, \theta^{(2)}) = \begin{vmatrix} \frac{2x_1}{\theta_1^{(1)^3}} & 0 \\ 0 & 0 \end{vmatrix} \quad \dots (5.25)$$



Therefore, we obtain

$$\rho(x) = \rho_1(x) = 1, \quad x \in \text{interior } A_5 \quad \dots \quad (5.26)$$

and it is easily seen that in fact (5.26) holds on interior  $A_i$  for  $i = 6, 7, 8$  as well. If  $x \in A_9$ , then  $p(x; \theta)$  reduces to a constant and hence all the mixed partial derivatives vanish; thus, for  $x \in \text{interior } A_9$  we have that

$$\rho(x) = \rho_1(x) = 0. \quad \dots \quad (5.27)$$

We may thus conclude that

$$R \supset \Omega - \bigcup_{i=5}^8 \text{bndry. } A_i.$$

On the other hand, an immediate calculation gives us that

$$\rho_1(x) < \rho(x) \quad x \in \bigcup_{i=5}^8 \text{bndry. } A_i \quad \dots \quad (5.28)$$

and therefore

$$R = \Omega - \bigcup_{i=5}^8 \text{bndry. } A_i. \quad \dots \quad (5.29)$$

That the Lebesgue measure of  $\Omega - R$  is 0 follows from the fact that  $\Omega - R$  is a linear set and thus has 0 planar measure.

Finally we define

$$T(x) = \begin{cases} (x_1, x_2) & x \in \bigcup_{i=1}^4 A_i \\ x_1^2 & x \in \text{interior } A_5 \cup \text{interior } A_8 \\ x_2^2 & x \in \text{interior } A_6 \cup \text{interior } A_7 \\ 0 & x \in \text{interior } A_9 \end{cases} \quad \dots \quad (5.30)$$

This defines  $T(x)$  a.e. (Lebesgue) on  $\Omega$  and we note that for  $x \in R$  with the proper choices of  $j_1, j_2, \theta^{(1)}$  and  $\theta^{(2)}$ , the  $T(x)$  defined by (5.30) is the statistic constructed, up to additive constants, in the proof of Theorem 3.3 and hence  $T(x)$  is a sufficient statistic for the family  $\mathcal{P}$  defined by (5.13) which is regular everywhere on  $R$  and of minimum dimension everywhere on  $R$ .



REFERENCES

- BAHADUR, R. R. (1954): Sufficiency and statistical decision functions. *Ann. Math. Stat.*, **25**, 423-462.
- BHATTACHARYYA, A. (1946): On some analogues of the amount of information and their use in statistical estimation, Chap. I. *Sankhyā*, **8**, 1-14.
- (1947): On some analogues of the amount of information and their use in statistical estimation, Chaps. II and III, *Sankhyā*, **8**, 201-218.
- (1948): On some analogues of the amount of information and their use in statistical estimation, Chap. IV, *Sankhyā*, **8**, 315-328.
- CARATHEODORY, C. (1935): *Variationsrechnung und Partielle Differentialgleichungen Erster Ordnung*, B. G. Teubner (Lithoprint Reproduction, 1945, J. W. Edwards).
- DARMOIS, G. (1935): Sur les lois de probabilité a estimation exhaustive. *C. R. de l'Acad. des Sc. de Paris*, **200**, 1265.
- (1937): Resumes exhaustifs d'un ensemble d'observations. *Bull. de l'Inst. Int. de Stat.*, **29**, 288-293.
- (1945): Sur les limites de la dispersion de certaines estimations. *Rev. de l'Inst. Int. de Stat.*, **13**, 9-15.
- DYNKIN, E. B. (1951): Necessary and sufficient statistics for a family of probability distributions. *Uspehi Matem. Nauk (N.S.)*, **6**, (1-41), 68-90.
- FISHER, R. A. (1922): On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London*, **222(A)**, 309-368.
- (1934): Two new properties of mathematical likelihood. *Proc. Roy. Soc. London*, **144(A)**, 285-307.
- HALMOS, P. R. AND SAVAGE, L. J. (1949): Application of the Randon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Stat.*, **20**, 225-241.
- KOOPMAN, B. O. (1936): On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.*, **39**, 399-409.
- LEHMANN, E. L. AND SCHEFFÉ, H. (1950): Completeness, similar regions and unbiased estimation. Part I., *Sankhyā*, **10**, 305-340.
- NEYMAN, J. (1935): Su un teorema concernente le cosiddette statistiche sufficienti. *Inst. Ital. Atti. Giorn.*, **6**, 320-334.
- RAO, C. R. (1945): Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calc. Math. Soc.*, **37**, 81-91.
- (1946): Minimum variance and the estimation of several parameters. *Proc. Camb. Phil. Soc.*, **43**, 280-283.
- (1948): Sufficient statistics and minimum variance estimates. *Proc. Camb. Phil. Soc.*, **45**, 213-218.
- (1952a): Some theorems on minimum variance estimation. *Sankhyā*, **12**, 27-42.
- (1952b): Minimum variance estimation in distributions admitting ancillary statistics. *Sankhyā*, **12**, 53-56.
- (1952c): On statistics with uniformly minimum variance. *Science and Culture*, **17**, 483-484.

*Paper received : June, 1958.*

# THE FAMILY OF ANCILLARY STATISTICS

By D. BASU

*Indian Statistical Institute, Calcutta*

**SUMMARY.** Though the marginal distributions of the ancillary statistics are independent of the parameter they are not useless or informationless. A set of ancillaries may sometimes summarise the whole of the information contained in the sample. A classification of the ancillaries in terms of the partial order of their information content is attempted here. In general there are many maximal ancillaries. Among the minimal ancillaries there exists a unique largest one. When there exists a complete sufficient statistic, the problem of tracking down the maximal and minimal ancillaries becomes greatly simplified.

## 1. INTRODUCTION

An ancillary<sup>1</sup> statistic is one whose distribution is the same for all possible values of the unknown parameter. A statistic that is not ancillary may be called 'informative'. The classical example of an ancillary statistic is the following:

*Example (a):* Let  $X$  and  $Y$  be two positive valued random variables with the joint density function

$$f(x, y) = e^{-\theta x - \frac{y}{\theta}}, \quad x > 0, y > 0, \theta > 0.$$

Here  $F = XY$  is an ancillary statistic. The maximum likelihood estimator  $T = \sqrt{Y/X}$  of  $\theta$  is not a sufficient statistic. However, the pair  $(F, T)$  is jointly sufficient.

The above example shows that though an ancillary statistic, by itself, fails to provide any information about the parameter, yet in conjunction with another statistic—which, as we shall presently see, need not be informative—may supply valuable information<sup>2</sup> about the parameter. In the following example we have given a family of ancillary statistics that are jointly equivalent to the whole sample.

*Example (b):* Let  $X$  and  $Y$  be independent normal variables with unknown means  $\theta$  and unit standard deviations. Here  $X - Y$  is an ancillary statistic. It is commonly believed that every ancillary statistic (in this situation) is necessarily a function of  $X - Y$ . That, however, is not true.

Let

$$F_c = F_c(X, Y) = \begin{cases} X - Y & \text{if } X + Y < c \\ Y - X & \text{if } X + Y \geq c \end{cases}$$

where  $c$  is a fixed constant.

---

<sup>1</sup> The name 'ancillary' is due to Fisher (1925). The name 'distribution-free' is also in use and perhaps would have been more appropriate in the present context.

<sup>2</sup> See Fisher (1956) for a discussion of how the ancillary information may (according to Fisher) be recovered.



Since  $X - Y$  and  $Y - X$  are identically distributed and each is independent of  $X + Y$  it at once follows that  $F_c$  is independent of  $X + Y$  and has the same distribution as that of  $X - Y$ . Thus,  $F_c$  is ancillary for each  $c$ . Consider now the family  $\{F_c\}$ ,  $-\infty < c < \infty$ , of ancillary statistics. For fixed  $X$  and  $Y$  the different values of  $F_c$  (for varying  $c$ ) are either  $X - Y$  or  $Y - X$ . The value  $c_0$  of  $c$  where  $F_c$  changes sign ( $F_c$  does not change sign only if  $X - Y = 0$  and that is a null event) is the value of  $X + Y$ . Thus, given  $F_c(X, Y)$  for all  $c$  we can find  $X + Y$  and  $X - Y$ . Hence, the family  $\{F_c\}$  of ancillary statistics is equivalent to the whole sample  $(X, Y)$ . The countable family  $\{F_c\}$  where  $c$  runs through the set of rational numbers is easily seen to be also equivalent to  $(X, Y)$ .

The author (Basu ; 1955, 1958) has shown that, under very mild restrictions, any statistic independent of a sufficient statistic is ancillary and that the converse proposition is also true, provided the sufficient statistic is complete.

In Example (b) the statistic  $T = X + Y$  is a complete sufficient statistic. A statistic  $F$  can, therefore, be ancillary if and only if  $F$  is independent of  $T$ . The following is a general method for constructing statistics independent of  $T$ . Start with any ancillary statistic  $F$ . In general, there will be many measure-preserving transformations of  $F$  (i.e. a mapping  $\varphi$  of the range space of  $F$  into itself such that  $\varphi(F)$  and  $F$  are identically distributed). For each real  $t$ , define a measure-preserving transformation  $\varphi_t$  of  $F$ . Then, take the statistic  $\varphi_t(F)$ . Subject to some measurability restrictions,  $\varphi_t(F)$  will be independent of  $T$  and hence will be ancillary. In Example (b) we took  $F = X - Y$  and  $\varphi_t(F) = F$  or  $-F$  according as  $t < c$  or  $\geq c$ .

If a statistic  $F$  is ancillary then every (measurable) function of  $F$  is also ancillary. The statistic  $F_2$  is said to include (or be more informative than) the statistic  $F_1$  if  $F_1$  can be expressed as a function of  $F_2$ . In this case we write  $F_2 \supset F_1$  or  $F_1 \subset F_2$ . Two statistics are said to be equivalent if each can be expressed as a function of the other.

*Example (c):* Let  $X_1, X_2, \dots, X_n$  be  $n$  independent observations on a normal variable with mean  $\theta$  and s.d. unity. Then each of the  $n-1$  statistics

$$F_1 = X_1 - X_2, F_2 = (X_1 - X_2, X_1 - X_3), \dots, F_{n-1} = (X_1 - X_2, X_1 - X_3, \dots, X_1 - X_n)$$

is ancillary and

$$F_1 \subset F_2 \subset \dots \subset F_{n-1}$$

The two ancillary statistics  $F_{n-1}$  and  $F = (X_2 - X_1, X_2 - X_3, \dots, X_2 - X_n)$  are easily seen to be equivalent.

From Example (b) it is obvious that  $F_{n-1}$  does not include all ancillary statistics.

An ancillary statistic  $M$  is said to be 'maximal' if there exists no non-equivalent ancillary  $M^*$  such that  $M \subset M^*$ . Thus, given any ancillary  $F$ , either it is maximal or there exists an ancillary  $F^* \supset F$ . Given any ancillary  $F_0$ , there exists (Theorem 2)

## THE FAMILY OF ANCILLARY STATISTICS

a maximal ancillary  $M \supset F_0$ . In general there exists many non-equivalent maximal ancillaries. A typical property (Cor. to Theorem 4) of a maximal ancillary  $M$  is that, for any ancillary  $F$  not included in  $M$ , the pair  $(M, F)$  is informative.

A minimal ancillary is one that is included in every maximal ancillary. Among the class of minimal ancillaries there exists (Theorem 5) a unique largest one  $G_0$ . In the absence of a better name we prefer to call  $G_0$  the laminal ancillary.  $G_0$  includes every minimal ancillary and is included in every maximal ancillary. A typical property (Theorem 6) of a minimal ancillary  $G$  is that, for any ancillary  $F$ , the pair  $(G, F)$  is ancillary.

If there exists a complete sufficient statistic  $G$ , then, any ancillary statistic  $F$ , such that the pair  $(G, F)$  is essentially equivalent to the whole sample, is shown (Theorem 7) to be essentially maximal. Under some further restrictions, the laminal ancillary is shown (Theorem 8) to be essentially equivalent to a constant.

In the following sections we elaborate on the above sketch of the family-tree of ancillary statistics. For the sake of elegance and brevity of exposition we use the language of sub  $\sigma$ -fields. Reference may be made to Bahadur (1954, 1955) for excellent expositions of the sub  $\sigma$ -field approach.

### 2. DEFINITIONS

Let  $(\mathcal{X}, \mathcal{B})$  be an arbitrary measurable space and let  $\{P_\theta\}$ ,  $\theta \in \Omega$  be a family of probability measures on  $\mathcal{B}$ . Any statistic  $T$  induces a sub  $\sigma$ -field  $\mathcal{B}_T \subset \mathcal{B}$ . Instead of dealing with statistics it is more convenient (in the present context) to deal with the corresponding sub  $\sigma$ -fields.

*Definition 1:* The event  $A \in \mathcal{B}$  is said to be ancillary if  $P_\theta(A)$  is the same for all  $\theta \in \Omega$ . The family of all ancillary events is denoted by  $\mathcal{A}$ .

It is easy to check that the family  $\mathcal{A}$  is closed for complementation and countable disjoint unions. However, in general  $\mathcal{A}$  is not closed for intersection (i.e.  $\mathcal{A}$  is not a  $\sigma$ -field).

In order to show that the family  $\mathcal{A}$  in Example (b) do not constitute a  $\sigma$ -field, we have only to check that

$$P_\theta [X - Y > 0 \text{ and } F_c(X, Y) > 0] = P_\theta (X - Y > 0 \text{ and } X + Y < c)$$

$$= \frac{1}{2} \int_{-\infty}^c \frac{1}{\sqrt{4\pi}} e^{-\frac{1}{4}(x-2\theta)^2} dx$$

which varies with  $\theta$ .

In Example (b) the Borel-extension of  $\mathcal{A}$  is  $\mathcal{B}$ .



*Example (d) :* Let  $\mathcal{X}$  consist of the three points  $a, b$  and  $c$  and let the corresponding probability measures be  $\frac{1}{4}-\theta, \frac{1}{2}$ , and  $\frac{1}{4}+\theta$  respectively, where  $0 < \theta < \frac{1}{4}$ . Here  $\mathcal{A}$  consists of the four sets  $\phi, [b], [a, c]$  and  $\mathcal{X}$  and so  $\mathcal{A}$  is a sub  $\sigma$ -field of  $\mathcal{B}$ .

*Definition 2 :* A  $\sigma$ -field  $\mathcal{F}$  is said to be ancillary if  $\mathcal{F} \subset \mathcal{A}$ . A  $\sigma$ -field that is not ancillary is called informative.

A statistic is ancillary or informative according as the corresponding  $\sigma$ -field is so.

*Definition 3 :* Two ancillary sets  $A$  and  $B$  are said to conform if  $AB$  is also ancillary. If  $A$  conforms to  $B$  then we write  $A \sim B$ . Since  $P_\theta(AB) + P_\theta(AB') = P_\theta(A)$  it follows that  $A \sim B$  if and only if  $A \sim B'$ .

If  $A$  conforms to every one of a sequence of disjoint sets  $B_1, B_2 \dots$  then it is easy to check that  $A \sim \bigcup B_i$ .

*Definition 4 :* Let  $\Gamma_0$  be the family of all ancillary sets  $B$  such that  $B \sim A$  for all  $A \in \mathcal{A}$ .

Clearly  $\phi$  and  $\mathcal{X}$  belong to  $\Gamma_0$ . From what we have said before it follows that  $\Gamma_0$  is closed for complementation and countable disjoint unions.

*Theorem 1 :* The family  $\Gamma_0$  is a  $\sigma$ -field.

*Proof :* It is enough to show that  $\Gamma_0$  is closed for intersection. Let  $B_1$  and  $B_2$  both belong to  $\Gamma_0$  and let  $A \in \mathcal{A}$ . From  $B_2 \in \Gamma_0$  it follows that  $B_2A \in \mathcal{A}$ . From  $B_1 \in \Gamma_0$  it then follows that  $B_1B_2A \in \mathcal{A}$ . Since  $A$  is an arbitrary ancillary set, it follows that  $B_1B_2 \in \Gamma_0$ .

We shall later on see that the ancillary  $\sigma$ -field  $\Gamma_0$  corresponds to the laminal ancillary  $G_0$  that we have referred to in §1.

The family  $\mathcal{A}$  of ancillary sets is a  $\sigma$ -field if and only if every pair of ancillary sets conform to one another, i.e. if  $\mathcal{A} = \Gamma_0$ .

*Example (e) :* Let  $\mathcal{X}$  consist of the five points  $a, b, c, d$  and  $e$  with the corresponding probabilities  $\frac{1}{2}, \theta, \theta, \frac{1}{4}-\theta$  and  $\frac{1}{4}-\theta$  respectively, where  $0 < \theta < \frac{1}{4}$ . In this case  $\Gamma_0$  consists of the four sets  $\phi, [a], [b, c, d, e]$ , and  $\mathcal{X}$ . The two sets  $[b, d]$  and  $[b, e]$  are both ancillary but they do not conform. Here  $\mathcal{A}$  is wider than  $\Gamma_0$  and is not a  $\sigma$ -field.

*Definition 5 :* The ancillary  $\sigma$ -field  $\mathcal{F}_2$  is said to include the ancillary  $\sigma$ -field  $\mathcal{F}_1$  (in symbols  $\mathcal{F}_2 \supset \mathcal{F}_1$  or  $\mathcal{F}_1 \subset \mathcal{F}_2$ ) if every element of  $\mathcal{F}_1$  is an element of  $\mathcal{F}_2$ .

The above partial order on ancillary  $\sigma$ -fields corresponds to the inclusion relationship for ancillary statistics.

*Definition 6 :* The ancillary  $\sigma$ -field  $\mathcal{M}$  is said to be maximal if there exists no other ancillary  $\sigma$ -field  $\mathcal{M}^*$  such that  $\mathcal{M}^* \supset \mathcal{M}$ .

## THE FAMILY OF ANCILLARY STATISTICS

*Definition 7 :* The intersection of all the maximal ancillary  $\sigma$ -fields is called the laminal ancillary.

The laminal ancillary is the largest ancillary that is included in all maximal ancillaries.

### 3. EXISTENCE AND CHARACTERIZATIONS OF MAXIMAL AND LAMINAL ANCILLARIES

The following theorem is fundamental.

*Theorem 2 :* Given any ancillary  $\sigma$ -field  $\mathcal{F}_0$  there exists a maximal ancillary  $\sigma$ -field  $\mathcal{M} \supset \mathcal{F}_0$ .

*Proof :* We first prove that given any family  $\{\mathcal{F}_j\}$ ,  $j \in J$  of ancillary  $\sigma$ -fields that are linearly ordered (by the inclusion relationship), the Borel-extension  $\mathcal{F}$  of  $\bigcup \mathcal{F}_j$  is also ancillary.

Clearly,  $\bigcup \mathcal{F}_j$  contains  $\phi$  and  $\mathcal{X}$  and is closed for complementation. Since  $\{\mathcal{F}_j\}$  is linearly ordered it follows that  $\bigcup \mathcal{F}_j$  is also closed for finite unions. That is,  $\bigcup \mathcal{F}_j$  is a field of sets.

Since each  $\mathcal{F}_j$  is ancillary, the restriction of  $P_\theta$  to  $\bigcup \mathcal{F}_j$  is a measure  $Q$  that does not depend on  $\theta$ . From the fundamental Extension Theorem of measures (Kolmogorov, 1933) we know that the extension of  $Q$  to  $\mathcal{F}$  is unique.

It follows at once that the restriction of  $P_\theta$  to  $\mathcal{F}$  is the same for all  $\theta$ , i.e.  $\mathcal{F}$  is an ancillary  $\sigma$ -field.

Now let  $\mathcal{C}$  be the family of all ancillary  $\sigma$ -fields that include  $\mathcal{F}_0$ . Since corresponding to any linearly ordered sub-family of  $\mathcal{C}$  there exists an ancillary  $\sigma$ -field that includes every member of the sub-family it follows from Zorn's Lemma that  $\mathcal{C}$  has a maximal element.

Let  $\{\mathcal{M}_i\}$ ,  $i \in I$  be the family of all maximal ancillary  $\sigma$ -fields. We at once have the

*Theorem 3 :*  $\mathcal{A} = \bigcup \mathcal{M}_i$

*Proof :* We have only to note that corresponding to any element  $A$  of  $\mathcal{A}$  there exists an ancillary  $\sigma$ -field that contains  $A$  as an element and then apply Theorem 2.

*Corollary :* If  $\{\mathcal{M}_i\}$  consists of only one  $\sigma$ -field  $\mathcal{M}_0$  then  $\mathcal{A} = \mathcal{M}_0 = \Gamma_0$ .

Thus, in any situation where there are non-conforming ancillary sets, the family  $\{\mathcal{M}_i\}$  has at least two members.

In Example (d) there is a unique maximal ancillary. In Example (e) there are exactly two maximal ancillaries namely :

$\mathcal{M}_1 =$  the  $\sigma$ -field spanned by  $[a]$  and  $[b, d]$

$\mathcal{M}_2 =$  the  $\sigma$ -field spanned by  $[a]$  and  $[b, e]$ .

and

*Theorem 4 :* If the ancillary set  $A$  does not belong to the maximal ancillary  $\mathcal{M}$  then  $A$  does not conform to at least one element of  $\mathcal{M}$ .



*Proof:* Suppose on the contrary that  $A$  conforms to every element of  $\mathcal{M}$ . Consider the family  $\mathcal{M}^*$  of sets  $AX \cup A'Y$  where  $X$  and  $Y$  are arbitrary elements of  $\mathcal{M}$ . Clearly  $\mathcal{M} \subset \mathcal{M}^*$  but not conversely.

Since  $(AX \cup A'Y)' = AX' \cup A'Y'$ ,

and  $\bigcup (AX_i \cup A'Y_i) = A (\bigcup X_i) \cup A' (\bigcup Y_i)$

and  $P_\theta (AX \cup A'Y) = P_\theta (AX) + P_\theta (A'Y)$ ,

it follows that  $\mathcal{M}^*$  is also an ancillary  $\sigma$ -field.

This, however, contradicts the maximality of  $\mathcal{M}$ .

*Corollary:* If  $\mathcal{M}$  be any maximal ancillary and if the ancillary  $\sigma$ -field  $\mathcal{F}$  is not included in  $\mathcal{M}$  then the smallest  $\sigma$ -field containing both  $\mathcal{M}$  and  $\mathcal{F}$  is informative.

*Theorem 5:*  $\bigcap \mathcal{M}_i = \Gamma_0$

*Proof:* Since every element of  $\Gamma_0$  conforms (by definition) to every ancillary event, it follows from Theorem 4 that  $\Gamma_0 \subset \mathcal{M}_i$  for all  $i$ , i.e.  $\Gamma_0 \subset (\bigcap \mathcal{M}_i)$ .

Now let  $B \in \bigcap \mathcal{M}_i$  and  $A$  be an arbitrary ancillary set. From Theorem 3 it follows that  $A \in \mathcal{M}_i$  for some  $i$ .

Hence  $B$  and  $A$  are together as elements of some  $\mathcal{M}_i$  and so  $B \sim A$ .

Since  $A$  is arbitrary it follows that  $B \in \Gamma_0$ .

$\therefore (\bigcap \mathcal{M}_i) \subset \Gamma_0$  and so the equality is proved.

*Theorem 6:* For any ancillary  $\sigma$ -field  $\mathcal{F}$  the smallest  $\sigma$ -field containing both  $\mathcal{F}$  and  $\Gamma_0$  is also ancillary.

*Proof:* Consider the family of 'rectangular' sets  $X \cap Y$  where  $X \in \mathcal{F}$  and  $Y \in \Gamma_0$ . From the definition of  $\Gamma_0$  it follows that all such sets are ancillary and that they conform to one another. The family of sets that may be formed by finite unions of rectangular sets form a field of sets and each of them is ancillary. The rest follows from the Extension Theorem of Measures.

#### 4. WHEN A COMPLETE SUFFICIENT STATISTIC EXISTS

In general there exist many maximal ancillaries. For instance, in Example (b) there are uncountably many maximal ancillaries. In order to see this, let us consider the family  $\{A_c\}$  of ancillary events where  $A_c = \{(X, Y) \mid F_c(X, Y) > 0\}$ . If  $c < d$ , then

$$P_\theta (A_c A_d) = P_\theta (X - Y > 0 \text{ and } X + Y < c) + P_\theta (Y - X > 0 \text{ and } X + Y \geq d)$$

$$= \frac{1}{2} \left[ 1 - \int_c^d \frac{1}{2\sqrt{\pi}} e^{-\frac{1}{4}(x-2\theta)^2} dx \right]$$

which varies with  $\theta$ .

Thus, the members of the family  $\{A_c\}$  of ancillary sets are mutually non-conforming. Hence the maximal ancillaries including the different members of the family are all different.

# THE FAMILY OF ANCILLARY STATISTICS

Though there may exist many maximal ancillaries, it is not, in general, easy to prove the maximality of a particular ancillary. However, in the situations where we have a complete sufficient statistic, it is rather easy to demonstrate the maximality of a large class of ancillaries.

The following property of complete sufficient statistics is useful.<sup>1</sup> Here we state and prove the result in terms of  $\sigma$ -fields.

**Lemma (Basu, 1955):** *If  $\mathcal{G} \subset \mathcal{B}$  be a boundedly complete sufficient  $\sigma$ -field and  $A$  any ancillary event, then  $A$  is independent of  $\mathcal{G}$ .*

*Proof:* Let  $\varphi = P(A|\mathcal{G})$  be the conditional probability of  $A$  given  $\mathcal{G}$ . That is,  $\varphi$  is a  $\mathcal{G}$ -measurable function such that

$$P_\theta(AG) = \int_G \varphi dP_\theta \text{ for all } \theta \in \Omega \text{ and } G \in \mathcal{G}.$$

Since  $\mathcal{G}$  is sufficient, it follows that  $\varphi$  may be chosen to be independent of  $\theta$ . Also the set of  $x$ 's for which  $\varphi(x)$  lies outside the interval  $(0, 1)$  is of zero-measure for each  $\theta \in \Omega$ .

Taking  $G = \mathcal{X}$  we have

$$P_\theta(A) = \int \varphi dP_\theta \text{ for all } \theta \in \Omega.$$

Since  $P_\theta(A)$  is independent of  $\theta$  and  $\varphi$  is  $\mathcal{G}$ -measurable, it follows from the bounded completeness of  $\mathcal{G}$  that  $\varphi = P_\theta(A)$  almost surely for all  $\theta \in \Omega$ .

$$\begin{aligned} P_\theta(AG) &= \int_G \varphi dP_\theta \\ \therefore &= P_\theta(A)P_\theta(G) \text{ for all } \theta \in \Omega \text{ and } G \in \mathcal{G}. \end{aligned}$$

That is,  $A$  is independent of all  $G \in \mathcal{G}$

Before proceeding further we need a slightly wider definition of maximality for an ancillary  $\sigma$ -field.

**Definition 8:** The two  $\mathcal{B}$ -measurable sets  $A$  and  $B$  are said to be essentially equal if

$$\begin{aligned} P_\theta(A \Delta B) &= P_\theta(AB' \cup A'B) \\ &= 0 \text{ for all } \theta \in \Omega. \end{aligned}$$

**Definition 9:** Two sub  $\sigma$ -fields  $\mathcal{F}_1$  and  $\mathcal{F}_2$  are said to be essentially equivalent if corresponding to any set belonging to one of them there exists an essentially equal set belonging to the other.

**Definition 10:** Any ancillary  $\sigma$ -field that is essentially equivalent to a maximal ancillary is called essentially maximal.

**Theorem 7:** *If  $\mathcal{G}$  be a boundedly complete sufficient  $\sigma$ -field then any ancillary  $\mathcal{F}$  such that the Borel-extension of  $\mathcal{G} \cup \mathcal{F}$  is essentially equivalent to  $\mathcal{B}$ , is essentially maximal.*

<sup>1</sup> See Basu (1955) and Hogg and Craig (1956) for some other interesting applications.



*Proof :* Let  $\mathcal{M}$  be a maximal ancillary including  $\mathcal{F}$  and let  $M$  be an arbitrary element of  $\mathcal{M}$ . For proving the essential maximality of  $\mathcal{F}$  we have to establish the existence of an  $F_0 \in \mathcal{F}$  such that  $F_0$  is essentially equal to  $M$ .

Let  $\mathcal{B}^*$  be the Borel extension of  $\mathcal{F} \cup \mathcal{G}$ . Since  $\mathcal{B}^*$  is essentially equivalent to  $\mathcal{B}$ , there exists an  $M^* \in \mathcal{B}^*$  such that  $M^*$  is essentially equal to  $M$ .

Since  $M \in \mathcal{M} \supset \mathcal{F}$  and  $M^*$  is essentially equal to  $M$ , it follows that  $M^*$  is an ancillary set conforming to every  $F \in \mathcal{F}$ . Clearly, the two measures  $P$  and  $Q$  on  $\mathcal{F}$ , defined by the relations  $P(F) = P_\theta(F)$  and  $Q(F) = P_\theta(M^*F)$ , are both independent of  $\theta$ .

Therefore, the conditional probability function

$$\varphi = P_\theta(M^*|\mathcal{F}) = \frac{dQ}{dP}$$

is independent of  $\theta$ .

Thus,  $\varphi$  is an  $\mathcal{F}$ -measurable function on  $\mathcal{X}$  such that

$$P_\theta(M^*F) = \int_F \varphi dP_\theta \text{ for all } \theta \in \Omega \text{ and } F \in \mathcal{F}.$$

Let  $F$  and  $G$  be typical elements of  $\mathcal{F}$  and  $\mathcal{G}$  respectively. Since  $\mathcal{F}$  is ancillary and  $\mathcal{G}$  is boundedly complete sufficient, it follows (from the Lemma) that  $\mathcal{F}$  and  $\mathcal{G}$  are independent.

$$\begin{aligned} \therefore \int_{FG} \varphi dP_\theta &= \int_{\mathcal{X}} (\varphi \mathcal{X}_F) \mathcal{X}_G dP_\theta \quad (\mathcal{X}_F \text{ and } \mathcal{X}_G \text{ are characteristic functions of } F \text{ and } G) \\ &= \int_{\mathcal{X}} (\varphi \mathcal{X}_F) dP_\theta \int_{\mathcal{X}} \mathcal{X}_G dP_\theta \quad (\because \mathcal{F} \text{ and } \mathcal{G} \text{ are independent}) \\ &= P_\theta(M^*F) P_\theta(G) \end{aligned} \quad \dots (\alpha)$$

Again, since  $M^* \sim F$  it follows (from the Lemma) that  $M^*F$  is independent of  $G$ .

$$\therefore \int_{FG} \mathcal{X}_{M^*} dP_\theta = P_\theta(M^*FG) = P_\theta(M^*F) P_\theta(G). \quad \dots (\beta)$$

From  $(\alpha)$  and  $(\beta)$  we have

$$\int_{FG} (\varphi - \mathcal{X}_{M^*}) dP_\theta = 0 \text{ for all } F \in \mathcal{F} \text{ and } G \in \mathcal{G}.$$

Since  $\varphi - \mathcal{X}_{M^*}$  is  $\mathcal{B}^*$ -measurable it at once follows that

$$\int_B (\varphi - \mathcal{X}_{M^*}) dP_\theta = 0 \text{ for all } B \in \mathcal{B}^*.$$

Therefore, for each  $\theta \in \Omega$ ,  $\varphi(x) - \mathcal{X}_{M^*}(x) = 0$  for almost all  $x$ .

Let  $F_0 = \{x | \varphi(x) = 1\}$ . Clearly  $F_0 \in \mathcal{F}$  and is essentially equal to  $M^*$ . Since  $M^*$  is essentially equal to  $M$  the Theorem is proved.

## THE FAMILY OF ANCILLARY STATISTICS

In Example (b),  $X + Y$  is a complete sufficient statistic. Also for any fixed  $c$ , the pair  $(X + Y, F_c)$  is equivalent to the sample  $(X, Y)$ . Hence it follows that every  $F_c$  is an essentially maximal ancillary. In Example (c), the ancillary  $F_{n-1}$  together with the complete sufficient statistic  $X_1 + X_2 + \dots + X_n$  is equivalent to the whole sample and, therefore, is essentially maximal. A large number of similar situations are covered by Theorem 7.

Having partially settled the question of maximal ancillaries let us turn our attention to the laminal ancillary.

The laminal ancillary is the largest ancillary  $\sigma$ -field that is included in all maximal ancillaries. From Theorem 5 we have that the class  $\Gamma_0$  of ancillary sets  $C$  that conform to every ancillary set is the laminal ancillary.

Let  $\Lambda$  be the family of sets that are essentially equal to either the empty set  $\phi$  or the whole space  $\mathcal{X}$ . That is,  $\Lambda$  is the family of all sets  $E$  such that  $P_\theta(E)$  is either  $\equiv 0$  or  $\equiv 1$  for all  $\theta \in \Omega$ . It is easy to check that  $\Lambda$  is a  $\sigma$ -field and that  $\Lambda \subset \Gamma_0$ . The following theorem covers a number of important cases.

**Theorem 8:** *If the following conditions are satisfied then  $\Gamma_0 = \Lambda$ .*

- i)  $\mathcal{F}$  is an essentially maximal ancillary.
- ii) There exists an informative set  $G$  which is independent of  $\mathcal{F}$ .
- iii) For every  $F \in \mathcal{F}$  such that  $0 < P_\theta(F) < 1$  there exists  $F^* \in \mathcal{F}$  such that  $P_\theta(F^*) = P_\theta(F)$  and  $P_\theta(FF^*) < P_\theta(F)$ .

*Proof:* Let  $C$  be an arbitrary element of  $\Gamma_0$ . We have to prove that  $P_\theta(C) = 0$  or  $1$ . If possible let  $0 < P_\theta(C) < 1$ .

Now,  $\mathcal{F}$  is essentially equivalent to a maximal ancillary and  $C$  belongs to every maximal ancillary. Hence, there exists  $F \in \mathcal{F}$  which is essentially equal to  $C$ . Thus,  $F$  conforms to every ancillary set and  $0 < P_\theta(F) < 1$ .

Let  $G$  and  $F^*$  satisfy conditions (ii) and (iii) respectively and let  $A = GF \cup G'F^*$ . Since  $G$  is independent of  $\mathcal{F}$ , we have

$$\begin{aligned} P_\theta(A) &= P_\theta(G)P_\theta(F) + P_\theta(G')P_\theta(F^*) \\ &= P_\theta(F)[P_\theta(G) + P_\theta(G')] \\ &= P_\theta(F). \end{aligned}$$

That is,  $A$  is an ancillary set.

Now

$$AF = GF \cup G'(FF^*)$$

and, therefore,

$$\begin{aligned} P_\theta(AF) &= P_\theta(G)P_\theta(F) + P_\theta(G')P_\theta(FF^*) \\ &= P_\theta(FF^*) + P_\theta(G)[P_\theta(F) - P_\theta(FF^*)]. \end{aligned}$$

Let us note that  $P_\theta(FF^*)$  and  $P_\theta(F) - P_\theta(FF^*)$  are both independent of  $\theta$  and that the latter is not zero. Again since  $G$  is informative  $P_\theta(G)$  is not independent of  $\theta$ . Hence  $AF$  is informative, which is a contradiction. Therefore,  $P_\theta(C) = 0$  or  $1$ , i.e.  $C \in \Lambda$ , which proves the theorem.



If the conditions of Theorem 7 are satisfied then  $\mathcal{F}$  and any informative  $G \in \mathcal{G}$  satisfies conditions (i) and (ii) of Theorem 8. We have then only to check whether condition (iii) is satisfied or not. If the restriction of  $P_\theta$  to  $\mathcal{F}$  be non-atomic then it is very easy to see that condition (iii) is also satisfied.

In Examples (b) and (c) the (essentially) maximal ancillaries have continuous (non-atomic) distributions and so Theorem 8 holds. Most of the familiar cases where a complete sufficient statistic exists fall under the above category.

*Example (f) :* Let  $X$  be a single observation on a normal variable with mean zero and standard deviation  $\sigma$ . Here  $X^2$  is a complete sufficient statistic.

$$\text{Let } Y = \begin{cases} -1 & \text{of } X < 0 \\ 1 & \text{if } X \geq 0 \end{cases}$$

Here the pair  $(Y, X^2)$  is equivalent to the whole sample  $X$ .

$\therefore Y$  is an essentially maximal ancillary.

The sub  $\sigma$ -field generated by  $Y$  consists of the four sets  $\phi$ ,  $(-\infty, 0)$ ,  $[0, \infty)$  and  $\mathcal{F}$ . Condition (iii) of Theorem 8 is clearly satisfied. Therefore, the laminal ancillary  $\Gamma_0$  is the same as  $\Lambda$ .

#### ACKNOWLEDGEMENT

I wish to thank Dr. R. R. Bahadur for some useful discussions.

#### REFERENCES

- BAHADUR, R. R. (1954): Sufficiency and statistical decision functions. *Ann. Math. Stat.*, **25**, 423.  
 ——— (1955): Statistics and subfields. *Ann. Math. Stat.*, **26**, 490.  
 BASU, D. (1955): On statistics independent of a complete sufficient statistics. *Sankhyā*, **15**, 377.  
 ——— (1958): On statistics independent of a sufficient statistic. *Sankhyā*, **20**, 223.  
 FISHER, R. A. (1925): Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700.  
 ——— (1956): *Statistical Methods and Scientific Inference*, Oliver and Boyd, London.  
 HOGG, R. V. and CRAIG, A. T. (1956): Sufficient statistics in elementary distribution theory. *Sankhyā*, **17**, 209.  
 KOLMOGOROV, A. N. (1933): *Foundations of The Theory of Probability*, Chelsea Publishing Company, New York.

*Paper received : July, 1959.*



# METRICIZING RANK-ORDERED OR UNORDERED DATA FOR A LINEAR FACTOR ANALYSIS

By LOUIS GUTTMAN<sup>1</sup>

*The Israel Institute of Applied Social Research and  
The Hebrew University, Jerusalem*

**SUMMARY.** Since the metrics of observed test scores are usually arbitrary, Thurstone posed the problem of how to "factor" them by using only rank-order considerations. One form of solution is to seek transformations that will yield new scores with a correlation matrix that is best from some point of view of factor analysis. But analyzing data via their correlation matrix is justified stochastically only if the regressions are linear. Assuming only the linearity restriction on regressions, it is shown that—in general—at most one set of new scores can be found to maintain the observed rank-orders. The factor-analyst has no freedom to mould the new correlation matrix by further considerations. More generally, it is shown how to compute *all* scoring systems which will yield linear regressions, the new scores possibly having polytone as well as monotone relations with the original rank order. In some cases, polytone transformations are the more appropriate ones. Similarly, computations are outlined for all ways of metricizing unordered data that will yield linear regressions.

## 1. INTRODUCTION

*The modified Thurstone problem of rank-order.* Thurstone (1947, pp. xiii-xiv) posed the problem of factor analysis of observed rank-orders in terms of having also the underlying factors only rank-ordered. A modification of this problem—which may in some cases be equivalent to Thurstone's more general statement—is as follows. For a set (i.e., population)  $P$  of subjects and a set  $J$  of tests, let  $s_{jp}$  be the observed score of subject  $p$  on test  $j$  ( $jp \in JP$ ).<sup>2</sup> The metrics of these observed scores are arbitrary up to monotone transformations within each test; only the rank-order of the  $s_{jp}$  for each  $j$  has fixed *a priori* meaning. Find those new scores  $x_{jp}$  ( $jp \in JP$ ) which will yield a correlation matrix possessing properties that are most desirable from some point of view of common-factor analysis, but which preserve the observed rank-orders within tests, or

$$\text{sign}(x_{jp} - x_{jq}) = \text{sign}(s_{jp} - s_{jq}) \quad (j p q \in JPP). \quad \dots (1.1)$$

Let  $r_{jk}$  denote the ordinary product-moment correlation coefficient between the desired new scores on tests  $j$  and  $k$  ( $jk \in JJ$ ), and let  $R$  denote the new correlation matrix,

$$R = [r_{jk}] \quad (jk \in JJ). \quad \dots (1.2)$$

<sup>1</sup> This research was facilitated by an uncommitted grant to the author from the Ford Foundation.

<sup>2</sup> By  $jp$  here is meant the *ordered pair* of elements  $j$  and  $p$ , or a *profile* over sets  $J$  and  $P$  in the indicated order. By  $JP$  is meant the Cartesian product of sets  $J$  and  $P$ , or the set of all possible profiles of the form  $jp$ , where  $j \in J$  and  $p \in P$ . This type of "facet" notation and terminology is convenient here and in other multivariate problems, making possible a compact but complete statement and analysis of a complex situation. To illustrate further,  $JPP$  in equation (1.1) below denotes a triple Cartesian product, or the set of all possible profiles of the form  $jpq$  where  $j \in J$ ,  $p \in P$ , and  $q \in P$ . While  $j$  and  $p$  are *elements* of  $J$  and  $P$  respectively, they are *components* of  $jp$ . Similarly,  $j$ ,  $p$ , and  $q$  are the *components* of profile  $jpq$ .  $J$  and  $P$  are called the *facets* of  $JP$  and  $JPP$ , the same  $P$  serving as two facets in the latter case.



The main diagonal elements of  $R$  are all unity, expressing complete self-correlations,

$$r_{jj} = 1 \quad (j \in J). \quad \dots (1.3)$$

Various criteria for  $R$  are possible. Thurstone himself, and many others, would prefer an  $R$  whose main diagonal could be modified to yield small rank for the reduced matrix (but keeping the Gramian property), to allow for specific and error factors (Thurstone, 1947). Others might seek an  $R$  parsimonious in a structural sense different from that of small rank, but also possibly modifying the main diagonal, (Guttman, 1954a, 1958a, 1958b). Still others would prefer not to modify the main diagonal, but to specify some structure for  $R$  as it is.

Previous investigators have tackled at least two different aspects of the rank-order factoring problem. Taking minimum rank for  $R$  as a desideratum, Bennett (1956) has developed an interesting algorithm for a lower bound to this minimum, in terms of absence of certain rank-order patterns among the observed scores. Since an actual  $R$  is his point of departure, Bennett's results refer to the modified Thurstone problem.

Using another point of departure,<sup>1</sup> Guttman (1946) derived heuristic equations for direct calculation of  $x_{jp}$  from the observed rank-orders, without pivoting on  $R$ , in terms of a smaller set of scores (principal components) which would tend to reproduce the observed rank-orders without assuming linear regressions. While the smaller set appears in metric form, actually only its rank-orders matter, and this solution may fit more closely into Thurstone's general formulation of the problem.

## 2. LINEARITY OF REGRESSION

It would seem, however, that in many cases, a linear factor analysis of an  $R$  would provide the best possible answer, especially if the *regressions were linear among the scores being factored*. In the case of such linear regressions, computation of the  $r_{jk}$  is stochastically justified, and  $R$  represents the actual interdependence of the new scores. Indeed, possible lack of linearity of regressions among the observed  $s_{jp}$  seems to be part of the original motivation for Thurstone to have posed his problem.

Even more restrictive stochastic conditions would be posited by Darmais' "general" factor analysis (Darmois, 1956) or by Lazarsfeld's (1950) latent structure theory.

In the present paper, an analysis of Thurstone's problem will be made, using only the stochastic condition of linearity of regression. This condition turns out to be so restrictive, that in general *at most one  $R$  can be found to satisfy it and (1.1) simultaneously*. This leaves no room for considering further desiderata of schools of

---

<sup>1</sup>The treatment (Guttman, 1946) is nominally for the case where judges do the ranking of objects; this is formally the same as our present problem if "judges" are interpreted to be our tests  $J$ , and "objects" are our subjects  $P$ .



# METRICIZING DATA FOR A LINEAR FACTOR ANALYSIS

factor analysis. The equations for computing the uniquely determined (up to linear transformations)  $x_{jp}$ —if they exist at all—are given in §§ 4-5 below. There is also the possibility that no satisfactory  $x_{jp}$  exist, or there is no solution to Thurstone's problem for a given set of data.

Solution of this version of Thurstone's problem is simplified by first considering and solving a more general problem. We actually study all possible transformations of the  $s_{jp}$ , polytone as well as monotone. Consequently, the  $s_{jp}$  need not be real numbers at all, nor even express rank-order. They may denote arbitrary qualitative categories.

The general problem solved in this paper is : Can real numbers be assigned to given qualitative categories for a given population  $P$  in such a way that the resulting numerical variables will have linear regressions on each other? If yes, in how many ways can this be done, and what are they? Should a positive answer be found for the given data, this may justify "factor analyzing" them via a product-moment correlation matrix. Otherwise, the qualitative or non-linear quantitative theories of scale analysis, latent structure analysis, facet analysis, etc., are called for.

In this broader context, Thurstone's problem is the special case where the categories are rank-orders, so that condition (1.1) can be considered as well.

## 3. NOTATION FOR GROUPED DATA

Let  $A_j$  be the set of all categories into which test  $j$  classifies members of  $P$  ( $j \in J$ ). A typical category (i.e., element) of  $A_j$  will be denoted by  $a$ , and sometimes by  $b$ . This "dummy" notation for categories requires reference to some set for meaning, and such reference will always be made. Let  $f_a$  be the proportion of  $P$  that falls into  $a$  ( $a \in A_j$ ,  $j \in J$ ). We shall consider only categories containing a positive proportion of  $P$ , or shall assume that

$$f_a > 0 \quad (a \in A_j, \quad j \in J). \quad \dots \quad (3.1)$$

We shall also assume that, for each  $j$ , the categories of  $A_j$  are mutually exclusive and exhaustive, so that

$$\sum_{a \in A_j} f_a = 1 \quad (j \in J). \quad \dots \quad (3.2)$$

Let  $m$  be the total number of tests, i.e., the number of elements of  $J$ . Let  $m_j$  be the number of categories in  $A_j$  ( $j \in J$ ). For Thurstone's problem, for each  $j$  the elements of  $A_j$  can be simply the integers from 1 to  $m_j$ , or—alternatively—the midpoints of  $m_j$  class-intervals for the  $s_{jp}$ . Tied ranks are explicitly allowed within a test, and  $m_j$  may vary from test to test—features which are quite prevalent in practice. More generally, no *a priori* order need exist among the element of  $A_j$  for any  $j$ ; they may be arbitrary qualities.



While we assume that  $m$  and each of the  $m_j$  are finite, no assumption will be made about the size of population  $P$ : it may be finite or infinite. Stochastically, the analysis makes most sense if  $P$  has an infinite number of members, for we shall not discuss sampling error due to selecting a subset of subjects from a larger population. Proportions such as  $f_a$  will be treated as if they had no sampling error. As far as the pure algebra goes, however,  $P$  below could be finite. The reader may make his own interpretation; this will not affect the formulae themselves.

To say that subject  $p$  has observed value  $s_{jp}$  on test  $j$  is to say that  $s_{jp} = a$ , where  $a$  is some category of  $A_j$ . Indeed, for fixed  $j$ ,  $f_a$  is simply the relative frequency over  $P$  with which the equality  $s_{jp} = a$  holds. Similarly, to say that subject  $p$  receives new score  $x_{jp}$  on test  $j$  is to say that, for this  $j$ , numerical value  $y_a$  is assigned to  $a$ , and  $x_{jp} = y_a$  whenever  $s_{jp} = a$ . To talk in terms of  $s_{jp}$  and  $x_{jp}$  is to talk about the "un-grouped" data, while categories  $a$  and scores  $y_a$  permit treatment of the data in "grouped" form.

A special case of "grouping" of course is where  $P$  is finite, with  $n$  members, and  $f_a \equiv 1/n$ , or each category of test  $j$  contains only one subject. This includes the case of untied rank-orders.

The correlation coefficients in  $R$  will not change if means and variances of the  $x_{jp}$  are changed. So there is no loss in generality in setting the means equal to 0 and the variances equal to 1. In grouped data form, this implies

$$\sum_{a \in A_j} f_a y_a = 0 \quad (j \in J) \quad \dots \quad (3.3)$$

and

$$\sum_{a \in A_j} f_a y_a^2 = 1 \quad (j \in J). \quad \dots \quad (3.4)$$

For fixed  $jk$ , let  $f_{ab}$  be the joint relative frequency of categories  $a$  and  $b$  ( $ab \in A_j A_k$ ). As usual, marginal frequencies are obtainable by summing over joint frequencies, or

$$f_a = \sum_{b \in A_k} f_{ab} \quad (a \in A_j, jk \in JJ). \quad \dots \quad (3.5)$$

By the usual product-moment formula for grouped data, recalling (3.3) and (3.4),

$$r_{jk} = \sum_{ab \in A_j A_k} y_a y_b f_{ab} \quad (jk \in JJ). \quad \dots \quad (3.6)$$

Clearly, if we consider two categories from the same test, or  $j = k$ ,

$$f_{ab} = \begin{cases} f_a & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases} \quad (ab \in A_j A_j, j \in J), \quad \dots \quad (3.7)$$



so the right member of (3.6) equals the left member of (3.4) when  $j = k$ , verifying (1.3).

Notice that for given  $jk$ , where  $j \neq k$ , (3.6) shows  $r_{jk}$  is a bilinear form over the  $f_{ab}$ .

#### 4. THE BIVARIATE CONTINGENCY MATRICES

For the bivariate case  $m = 2$ , the basic formulae and results to be presented now (in §§ 4-6) have been rediscovered independently by various writers during the past two decades<sup>1</sup> (cf. Bennett, 1956; Burt, 1950; Guttman, 1941, 1953a and 1953b; Hirschfeld, 1935; Maung, 1942; Williams, 1952). Despite their relative simplicity and their importance in a wide variety of situations, such formulae have not yet attracted the general attention they seem to deserve. Our present task is merely to extend them to the multivariate case. But even when  $m > 2$ , the bivariate regressions must be linear as well as the multiple regressions. A great deal of our work is accomplished by considering first all the bivariate regressions among the  $m$  tests.

For each pair of tests  $jk$ , the  $f_{ab}$  represent a contingency table, or matrix, of  $m_j$  rows and  $m_k$  columns. The elements of  $A_j$  are the row captions, and the elements of  $A_k$  are the column captions. The problem is to replace these captions by real numbers,  $y_a$  and  $y_b$  ( $ab \in A_j A_k$ ), so that the resulting numerical regressions will be linear.

Linearity of regression of the  $x_{jp}$  on the  $x_{kp}$ , say, is defined by considering, for each  $b$  in turn ( $b \in A_k$ ), the arithmetic mean of the  $x_{jp}$  for all  $p$  whose  $x_{kp}$  equals  $y_b$ . These conditional means must lie on a straight line when the  $y_b$  are abscissas, with slope  $r_{jk}$ . (The slope is more generally the regression coefficient, but this equals  $r_{jk}$  when variances are set equal to unity, as we have assumed). Expressing the regression of the  $x_{jp}$  on the  $x_{kp}$  in grouped data form, the linearity condition is

$$r_{jk}y_b = \frac{1}{f_b} \sum_{a \in A_j} y_a f_{ab} \quad (b \in A_k, jk \in JJ). \quad \dots (4.1)$$

The left member of (4.1) is the regression estimate of  $x_{jp}$  as a linear function of  $x_{kp} = y_b$ , while the right member is the direct statement of the conditional mean of the  $x_{jp}$  for fixed  $b$  (or  $y_b$ ).

The converse condition for linearity of regression, for that of the  $x_{kp}$  on the  $x_{jp}$ , is—analogueous to (4.1)

$$r_{jk}y_a = \frac{1}{f_a} \sum_{b \in A_k} y_b f_{ab} \quad (a \in A_j, jk \in JJ). \quad \dots (4.2)$$

Because of the "dummy" nature of the notation, (4.2) is strictly equivalent to (4.1). However, it is convenient to be able to refer to both forms.

---

<sup>1</sup> I am indebted to Dr. William Kruskal for supplying me with three of these references.



## 5. THE NUMBER OF SOLUTIONS IN THE BIVARIATE CASE

Multiplying (4.1) through by  $f_b y_b$ , summing over  $b \in A_k$ , and recalling (3.4) serve to verify (3.6). Similar verification comes from multiplying (4.2) through by  $f_a y_a$  and summing over  $a \in A_j$ . Indeed, for fixed  $jk$ , (4.1) and (4.2) are the *stationary equations for determining the  $y_a$  and  $y_b$  that will maximize (minimize)  $r_{jk}$  in (3.6), subject to restraint (3.4)*. This has been discovered by several of the writers cited above, by differentiating the right member of (3.6) subject to condition (3.4). Further known theorems are as follows.

For fixed  $jk$ , the number of linearly independent solutions to (4.1) and (4.2) is always equal to the rank of the matrix  $[f_{ab}]$ . All but one of these solutions will also satisfy (3.3). The one improper solution is always  $y_a \equiv y_b \equiv 1$ , for which  $r_{jk} = 1$ . Thus, the number of linearly independent *proper* solutions is one less than the rank of  $[f_{ab}]$ . To eliminate the one improper solution, define  $f'_{ab}$  by

$$f'_{ab} = f_{ab} - f_a f_b \quad (ab \in A_j A_k, jk \in JJ), \quad \dots (5.1)$$

and use  $f'_{ab}$  in place of  $f_{ab}$  in (4.1) and (4.2). The rank of  $[f'_{ab}]$  is one less than the rank of  $[f_{ab}]$ , and every solution of the new equations will be a solution of the old equations, and will also satisfy (3.3).  $f'_{ab}$  is simply  $f_{ab}$  with "chance expectation" removed as in computations for the chi-square test of significance for a contingency table.

## 6. RELATION TO CHI-SQUARE

There is an intimate relation with the entire chi-square theory. For fixed  $jk$ , let  $\rho_{jk}$  be the rank of  $[f'_{ab}]$ , so that  $\rho_{jk} + 1$  is the rank of  $[f_{ab}]$ . Let  $y_a^\rho$  and  $y_b^\rho$  be the  $\rho$ -th proper solution to (4.1) and (4.2), with resulting correlation  $r_{jk}^\rho$  ( $\rho = 1, 2, \dots, \rho_{jk}$ ). Then  $r_{jk}^\rho \neq 0$  always, and

$$\sum_{\rho=1}^{\rho_{jk}} (r_{jk}^\rho)^2 = \gamma_{jk}^2 \leq 1 \quad (jk \in JJ), \quad \dots (6.1)$$

where  $\gamma_{jk}^2$  is Karl Pearson's mean square contingency coefficient,

$$\gamma_{jk}^2 = \sum_{ab \in A_j A_k} \frac{(f_{ab} - f_a f_b)^2}{f_a f_b} \quad (jk \in JJ). \quad \dots (6.2)$$

If  $P$  were finite, say with  $n$  members, and if  $\chi_{jk}^2$  were computed in the usual fashion for the contingency matrix  $[f_{ab}]$  between tests  $j$  and  $k$ , then  $\chi_{jk}^2 = n\gamma_{jk}^2$  ( $jk \in JJ$ ). This helps explain why  $\chi_{jk}^2$  is often not a very sensitive test for statistical independence; it crudely averages contributions of  $\rho_{jk}$  possibly different sources of dependence without regard to the possible *structure* of dependence (Guttman, 1953b and Williams, 1952), that is, without specifying alternatives to the null hypothesis of independence as is required, say, in the Neyman-Pearson theory of testing hypotheses.



# METRICIZING DATA FOR A LINEAR FACTOR ANALYSIS

While the different solutions may not actually be statistically independent of each other, and hence not necessarily constitute independent sources of variation, they are always mutually *orthogonal*, or *linearly uncorrelated* if their latent roots are all unequal. More explicitly, for fixed  $jk$ , let  $y_a^\rho$  and  $y_a^{\rho'}$  be the respective scores given to a ( $a \in A_j$ ) by solutions  $\rho$  and  $\rho'$ , yielding correlations  $r_{jk}^\rho$  and  $r_{jk}^{\rho'}$  respectively. If these two correlation coefficients are unequal in absolute value—which is the general case when  $\rho \neq \rho'$ —then

$$\sum_{a \in A_j} f_a y_a^\rho y_a^{\rho'} = 0 \quad (\rho \neq \rho', j \in J). \quad \dots (6.3)$$

The above known facts will now be used to obtain new results, relevant to our present problems.

## 7. NUMBER OF SOLUTIONS MAINTAINING THE OBSERVED RANK-ORDER

How to solve (4.1) and (4.2) for fixed  $jk$  is well-known. For each  $\rho$ , an iterative procedure can be used to go from (4.1) to (4.2) and back again; or else (4.1) can be substituted in (4.2) to eliminate the  $y_b$ , and then iterations can be performed for the  $y_a$  alone from the resulting equations.

Now,  $\rho_{jk} + 1$  cannot exceed the smaller of  $m_j$  or  $m_k$ , since the rank of a matrix cannot exceed either its row or its column order. So the smaller of  $m_j - 1$  and  $m_k - 1$  is an upper bound to the number of solutions to the bivariate case. In the multivariate case, let  $m_0$  be the smallest of the  $m_j$  ( $j \in J$ ). Then  $m_0 - 1$  is an upper bound to the number of different ways in which regressions of the  $s_{jp}$  can be linearized.

If  $m_0 = 2$ , or at least one of the tests provides only a dichotomous classification for  $P$ , then there is at most one solution to Thurstone's problem, as well as to the more general problem of metricizing unordered qualitative data. Only if  $m_0 > 2$  is there room for more than one solution.

If  $m_0 > 2$  and  $m = 2$ ,  $\rho$  can possibly take on more than one value; nevertheless, Thurstone's problem again can have but one solution in general. In general, at most one of the linearized regression systems will satisfy (1.1). To prove this, again let  $y_a^\rho$  and  $y_a^{\rho'}$  be the respective scores given to a ( $a \in A_j$ ) by solutions  $\rho$  and  $\rho'$ , and let  $\theta_j$  be defined by

$$\theta_j = \sum_{a, b \in A_j} f_a f_b (y_a^\rho - y_b^\rho)(y_a^{\rho'} - y_b^{\rho'}) \quad (j \in J). \quad \dots (7.1)$$

Notice that both  $a$  and  $b$  in (7.1) are elements of the same  $A_j$ .

Now, if  $a$  and  $b$  are distinct categories of  $A_j$  that maintain the original rank-order (1.1), whether scored by  $\rho$  or by  $\rho'$ , then

$$(y_a^\rho - y_b^\rho)(y_a^{\rho'} - y_b^{\rho'}) > 0 \quad (a \neq b, a, b \in A_j, j \in J). \quad \dots (7.2)$$



From (3.1), (7.2), and (7.1), we obtain

$$\theta_j > 0 \quad (j \in J). \quad \dots (7.3)$$

But expanding the right member of (7.1) with the aid of (3.2), (3.3), and (6.3) yields, in general,

$$\theta_j = 0 \quad (j \in J), \quad \dots (7.4)$$

which contradicts (7.3). Therefore (7.2) cannot hold in general. No two distinct solutions can maintain the same rank-order (apart from the exceptional case of equal latent roots) for categories of the same test.

If there is a regression linearizing solution to Thurstone's problem, it is unique in general. In particular, *if the original  $s_{jp}$  already have linear regressions, there is no profit in general in trying to transform them into new scores* if condition (1.1) is to be satisfied.

Interesting empirical examples have been reported in (Guttman, 1953b), where condition (1.1) was deliberately abandoned to obtain psychologically more appropriate polytone transformations of rank-orders. For fixed  $jk$ , these linearized the regressions to maximize  $r_{jk}$ . This type of situation may prove to be fairly frequent with attitudinal data, should investigators begin to look into their results for possible polytone properties.

#### 8. THE MULTIVARIATE CASE

Thus far, we have focussed on solutions to (4.1) and (4.2) for fixed  $jk$ . But there is no guarantee that if scores  $y_a$  ( $a \in A_j$ ) are a solution for the joint distribution of test  $j$  with test  $k$ , they will remain a solution for test  $j$  with test  $i$  where  $k \neq i$ . For the case  $m > 2$ , the  $y_a$  ( $a \in A_j$ ) must linearize regressions *simultaneously* with all other tests if they are to be part of a solution for the multivariate case. If no scores  $y_a$  ( $a \in A_j$ ) do such a job with all tests, then there is no solution at all to the regression linearizing problem, whether or not condition (1.1) is to be considered, and whether or not the data are initially unordered.

A way of testing the simultaneous linearizing property is to compute the  $y_a$  and  $y_b$  ( $ab \in A_j A_j$ ) for some fixed  $jk$ , and then compute  $y_c$  ( $c \in A_i$ ) from each of the fixed scores by using the latter in the right of (4.1) and (4.2) respectively, but writing  $c$  and  $r_{ji}$  in place of  $b$  and  $r_{jk}$  in (4.1) and writing  $c$  and  $r_{ki}$  in place of  $a$  and  $r_{jk}$  in (4.2). The resulting left members should be numerically equal for all  $i \in J$ .

The above is also a way of avoiding repeatedly to solve (4.1) and (4.2) for various  $jk$ . Once solved for fixed  $jk$ , solutions for other pairs of tests can be got directly from these if the *simultaneous* linearizing property holds.

But even all this concerns only bivariate regressions, and gives no assurance about linearity of *multiple* regressions. For example, if all test scores were dichotomous ( $m_j = 2, j \in J$ ), simultaneous linearizing *always* holds, since a straight line can always be fitted to two points; there is always one and only one  $R$  (apart from



# METRICIZING DATA FOR A LINEAR FACTOR ANALYSIS

reflections of sign) that is proper from the bivariate distributions, namely that of point correlations from fourfold tables. But this  $R$  says little or nothing in general about the shape of the *multiple* regressions among the dichotomies.

Study of the multiple regressions cannot, of course, introduce new solutions beyond those obtained by the bivariate considerations of (4.1) and (4.2). What it can do is to eliminate some or all of the latter. For example, the unique solutions to (4.1) and (4.2) for dichotomies need not yield linear multiple regressions. In such a case, a "factor analysis" of the  $R$  of point correlations is inappropriate; nonlinear techniques are needed that go beyond  $R$ .

## 9. FORMULAE FOR THE MULTIPLE REGRESSIONS

To state the multiple regression conditions the  $y_a$  ( $a \in A_j$ ,  $j \in J$ ) must satisfy beyond (4.1) and (4.2); more notation is needed.

Let  $J_{jk}$  be the subset of  $J$  defined by omitting tests  $j$  and  $k$ , where  $j \neq k$ . Let  $C_{jk}$  be the Cartesian product of the sets of categories of the tests in  $J_{jk}$ ,

$$C_{jk} = \prod_{i \in J_{jk}} A_i \quad (j \neq k, jk \in JJ), \quad \dots \quad (9.1)$$

and let  $c$  be a typical element of  $C_{jk}$ , i.e., a profile over all tests except  $j$  and  $k$ . Let  $f_{abc}$  be the proportion of joint occurrences of  $abc$  ( $abc \in A_j A_k C_{jk}$ ,  $jk \in JJ$ ). Then the bivariate contingencies are given by

$$f_{ab} = \sum_{c \in C_{jk}} f_{abc} \quad (ab \in A_j A_k, jk \in JJ). \quad \dots \quad (9.2)$$

Similarly, if  $f_{bc}$  is the proportion of joint occurrences of  $b$  with  $c$ , then

$$f_{bc} = \sum_{a \in A_j} f_{abc} \quad (bc \in A_k C_{jk}, jk \in JJ). \quad \dots \quad (9.3)$$

Let  $y_{bc}$  be the predicted value of  $y_a$  on the  $j$ -th test for a subject whose profile on the  $m-1$  remaining tests is  $bc$ , according to some metricization of the original scores ( $abc \in A_j A_k C_{jk}$ ). Then for those profiles for which the members of (9.3) do not vanish,

$$y_{bc} = \frac{1}{f_{bc}} \sum_{a \in A_j} y_a f_{abc} \quad (bc \in A_k C_{jk}, jk \in JJ), \quad \dots \quad (9.4)$$

or  $y_{bc}$  is the conditional mean of the  $y_a$  for the given profile. Equality (9.4) defines the true regressions, whether or not they are linear.

If  $y'_{bc}$  denotes the estimate of  $y_a$  as a linear function of the scores on the remaining  $m-1$  tests ( $abc \in A_j A_k C_{jk}$ ), it can be expressed as follows. Let

$$\delta_{abc} = \begin{cases} 1 & \text{if } a \text{ is a component of } bc \\ 0 & \text{otherwise} \end{cases} \quad (abc \in A_i A_k C_{jk}, ijk \in JJJ). \quad \dots \quad (9.5)$$



Clearly, if  $i = j$  or  $j = k$ ,  $\delta_{abc} = 0$  in (9.5) for  $c_{jk}$  has no component from test  $j$  by definition, and is not defined if  $j = k$ . Let  $w_{jk}$  be the multiple regression coefficient of test  $k$  in the linear regression of test  $j$  on the remaining  $m-1$  tests ( $jk \in JJ$ ). As usual,  $w_{jj} \equiv 0$ , for a test is not used to predict itself. The  $w_{jk}$  can be computed directly from  $R$ , and most simply from  $R^{-1}$  when  $R$  is non-singular. If  $R$  is singular, the  $w_{jk}$  are not uniquely determined, but—as is well known—the predictions themselves are invariant with respect to choice of equally good  $w_{jk}$ , and can be stated as

$$y'_{bc} = \sum_{i \in J} \sum_{a \in A_i} y_a \delta_{abc} w_{ji} \quad (bc \in A_k C_{jk}, jk \in JJ). \quad \dots (9.6)$$

In the right member of (9.6), the summation over  $a \in A_i$  picks out the scores of the predicting tests which are associated with the profile  $bc$ , and then the summation over  $i \in J$  simply weights these by the regression coefficients and adds the products.

The linearity requirement for multiple regression is that  $y_{bc} \equiv y'_{bc}$ , or from (9.4) and (9.6),

$$\sum_{a \in A_j} y_a f_{abc} = f_{bc} \sum_{i \in J} \sum_{a \in A_i} y_a \delta_{abc} w_{ji} \quad (bc \in A_k C_{jk}, jk \in JJ). \quad \dots (9.7)$$

In the form (9.7), we need not worry about the vanishing of members of (9.3), for then both members of (9.7) vanish by virtue of (9.7) and non-negativeness of proportions.

To check (9.7) with empirical data may usually be prohibitive. As a partial check which may be serviceable in practice, one can use the following necessary condition derived from (9.7) and (4.1).

$$\text{Let} \quad g_{jab} = \sum_{c \in C_{jk}} f_{bc} \delta_{abc} \quad (ab \in A_i A_k, ijk \in JJJ). \quad \dots (9.8)$$

Then summing (9.7) over  $c \in C_{jk}$  and using (9.2), (4.1), and (9.8) yield

$$r_{jk} f_{jb} y_b = \sum_{i \in J} \sum_{a \in A_i} y_a g_{jab} w_{ji} \quad (b \in A_k, jk \in JJ). \quad \dots (9.9)$$

To use (9.9), first employ (4.1) and (4.2) for determining the  $y_a$  for each variable, compute the resulting  $R$  and  $w_{ji}$ , tabulate the  $g_{jab}$ , and see if all these hang together as required by (9.9). If (9.9) is not satisfied, the multiple regressions are not linear for this metricization.

Even weaker, but computationally more feasible, necessary conditions can be derived by summing the respective members of (9.9) over  $j$  or in other manners.

#### 10. RELATION TO OTHER STOCHASTIC CONSIDERATIONS

The importance of considering the nature of the true regressions when computing product-moment correlations is illustrated by the fact that zero correlation does not necessarily imply statistical independence. Consider a symmetric,  $U$ -shaped perfect regression of one variable on another: the correlation ratio is 1 while the linear correlation coefficient is 0. The perfect dependence of one variable on the other is obscured by the linear analysis.



## METRICIZING DATA FOR A LINEAR FACTOR ANALYSIS

While many—if not most—of the theorems of the various forms of linear factor analysis are non-statistical, general theorems for abstract Euclidean vector spaces, their use in the behavioral sciences requires stochastic interpretation. The substantive interest is in actual interdependencies, not in linear algebra.

To what extent is interdependence expressed by regression equations, linear or otherwise? If correlation ratios were used throughout instead of correlation coefficients, would these always tell the whole story? The answer to the second question is: No. A zero correlation ratio only states that conditional means of one variable on another do not vary. But while the means are constant, dependence may occur in the form of heteroscedasticity, varying skewnesses, etc.

A complete analysis should account for all forms of dependence. This is the motivation of both Darmon (1956) and Lazarsfeld (1950) in their respective approaches.

Our present analysis shows that the criterion of linear regressions is so restrictive by itself that it is not very hopeful that other criteria could often be satisfied as well in metricizing data. The multivariate normal distribution is exceptional in that it is so well behaved: zero linear correlation implies complete statistical independence.

If metricizing cannot lead to a multivariate normal distribution, the next best thing for many purposes is to obtain at least a well-defined regression system. How to seek the simplest linear system has been the topic of this paper.

Some writers seem to imply that if the  $s_{jp}$  are transformed so as to make their marginal distributions normal, then the new bivariate or multivariate distributions will be normal. Unfortunately, this is not necessarily the case, for marginal transformations can often say little about linearity or non-linearity of the resulting regressions. This is one of the possible fallacies in using the tetrachoric coefficient (cf. Guttman, 1950a). If marginal transformations could do the trick, Thurstone's problem would not have arisen, and the algebra above would be necessary.

From another point of view, our analysis does account completely for the dependence of the tests on each other. For fixed  $jk$ , if we consider all of the  $\rho_{jk}$  solutions to (4.1) and (4.2) and not select merely one of them—they completely account for the  $f_{ab}$ , for as several of our references have essentially shown (cf. Hirshfeld, 1935; Guttman, 1941, 1950b; Maung, 1942; Williams, 1952) it is always true that

$$\frac{f_{ab}}{f_a f_b} = 1 + \sum_{\rho=1}^{\rho_{jk}} r_{jk}^{\rho} y_a^{\rho} y_b^{\rho} \quad (ab \in A_j A_k, jk \in JJ). \quad \dots \quad (10.1)$$

The complete sets of regression linearizing scores always completely reproduce the observed pairwise occurrences, or completely "explain" bivariate dependence. To have multivariate dependence also accounted for this way would require that the pairwise solutions for  $jk$  also hold for  $ij$  and  $ik$  ( $i \neq j, k$ ), etc. This last condition is of course only necessary, and not sufficient.



The theorem of § 7 above states that in general no two metricizations of the categories of a test can maintain the same rank-order for given  $j$ . This means that  $y_a^\rho$  are in general a polytone function of the  $y_a^{\rho'} (a \in A_j)$  when  $\rho \neq \rho'$ , and conversely. To have but one of the  $\rho_{jk}$  solutions be basic in accounting for dependence would require all the others to be orderly polytone functions of it. An excellent example of a law of formation of polytone dependence which singles out one solution as basic occurs in the theory of perfect scales (Guttman, 1950b, 1954b).

This paradox of polytone relations among regression linearizing solutions seems worth exploring in more complex situations of rank-order. It has an important bearing not only on the chi-square theory of statistical dependence, but also on the structural problems with which factor analysis is concerned.

## REFERENCES

- BENNETT, J. F. (1956): Determination of the number of independent parameters of a score matrix from the examination of rank orders. *Psychometrika*, **21**, 383-393.
- BURT, C. (1950): The factorial analysis of qualitative data. *Brit. J. Psychol. Stat. Sect.*, **3**, 166-185.
- DARMOIS, G. (1956): Observations théoriques sur l'analyse factorielle, linéaire et générale. *L'analyse factorielle et ses applications*, Centre National de la Recherche Scientifique, Paris.
- FISHER, R. A. (1955): *Statistical Methods for Research Workers*, 12-th Edition. Oliver and Boyd, London, esp. pp. 289-295.
- GUTTMAN, L. (1941): The quantification of a class of attributes. In P. Horst, et al., *The Prediction of Personal Adjustment*, pp. 321-347. Social Science Research Council, New York.
- (1946): An approach for quantifying paired comparisons and rank order. *Ann. Math. Stat.*, **17**, 144-163.
- (1950a): Relation of scalogram analysis to other techniques. In S. A. Stouffer, et al., *Measurement and Prediction*; chap. 6. Princeton University Press.
- (1950b): The principal components of scale analysis. In Stouffer, *ibid.*, chap. 9.
- (1953a): A note on Sir Cyril Burt's 'Factorial analysis of qualitative data'. *Brit. J. Psychol. Stat. Sect.*, **6**, 1-4.
- (1953b): Two new approaches to factor analysis. Technical report on project Nonr-731(00), Office of Naval Research, Washington, D. C., Esp. Part II.
- (1954a): A new approach to factor analysis: the radex. In P. F. Lazarsfeld (ed.) *Mathematical Thinking in the Social Sciences*, chap. 6. The Free Press: Glencoe, Illinois.
- (1945b): The principal components of scalable attitudes. In Lazarsfeld, *ibid.* 5.
- (1958a): To what extent can communalities reduce rank? *Psychometrika*, **23**, 297-308.
- (1958b): What lies ahead for factor analysis? *Educational and Psychological Measurement*, **18**, 497-515.
- HIRSCHFELD, H. O. (1935): A connection between correlation and contingency. *Proc. Camb. Phil. Soc.*, **31**, 520-524.
- LAZARSFELD, P. F. (1950): The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, et al., *Measurement and Prediction*, chap. 10. Princeton University Press.
- MAUNG, K. (1942): Measurement of association in a contingency table, with special reference to the pigmentation of hair and eye colours of Scottish school children. *Ann. Eugenics*, **11**, 189-205.
- THURSTONE L. L. (1947): *Multiple-Factor Analysis*. University of Chicago Press.
- WILLIAMS, E. J. (1952): Use of scores for the analysis of association in contingency tables. *Biometrika*, **39**, 274-289.

Paper received: January, 1958.



# POSITIVE AND NEGATIVE DEPENDENCE OF TWO RANDOM VARIABLES

By H. S. KONIJN

*University of Sydney, Australia*

**SUMMARY.** Two random variables will be called completely positively dependent if there is an almost sure nondecreasing relation between them, and positively  $\kappa$ -dependent if their joint distribution is a mixture (with mixture coefficient  $\kappa$ ) of the distribution of two completely positively dependent random variables and the distribution of two independent random variables with the same marginals. Similar definitions can be given for complete negative dependence and negative  $\kappa$ -dependence. The paper discusses properties of such variates and properties of the power of various tests for independence against such types of dependence.

## 1. INTRODUCTION

In another paper (Konijn, 1956), the author has investigated the asymptotic power of various tests for independence of two random variables against the alternative that their joint distribution has been obtained by a nontrivial linear transformation of independent random variables and has given some justification for considering such a type of alternatives. Among the many other types one could consider, it seems natural to examine one under which the joint distribution belongs to a family of bivariate distributions, all with the same marginals, which, as an extreme, contains one implying a monotone relation between the variates. In fact, for any two given marginals, there is a unique joint distribution which implies a nondecreasing relation between the variates, and a unique joint distribution which implies a nonincreasing relation. We shall use what is mathematically perhaps the simplest method of constructing such families of distributions, by linearly combining such an extreme and the distribution corresponding to independence with the same marginals.

It is true that the bivariate distributions belonging to such families (other than those corresponding to the case of independence) all have a singular component, which seems rather unusual. Nevertheless, it may well be that a number of situations of practical interest can be represented approximately by a distribution of this family with the singular component replaced by a narrow band of probability mass about the corresponding curve in the plane of the two variates.

## 2. DEFINITION AND CHIEF PROPERTIES OF $F_+$ AND $F_-$

If  $Y_0$  and  $Z_0$  are two independent random variables with distributions  $G$  and  $H$ , we shall denote the distribution of  $X_0 = (Y_0, Z_0)$  by  $F_0 = GH$ . By  $F_+[F_-]$  we shall denote the uniformly highest [lowest] valued bivariate distribution function with marginals  $G$  and  $H$ ; that these exist and are given by

$$F_+(y, z) = \begin{cases} G(y) & \text{if } G(y) \leq H(z), \\ H(z) & \text{if } G(y) \geq H(z) \\ 0 & \text{if } G(y) + H(z) - 1 \leq 0 \end{cases}$$

$$F_-(y, z) = \begin{cases} 0 & \text{if } G(y) + H(z) - 1 \leq 0 \\ G(y) + H(z) - 1 & \text{if } G(y) + H(z) - 1 \geq 0 \end{cases}$$



was shown by Fréchet (1951). In fact, these expressions are easily seen to bound any bivariate distribution function with given marginals  $G$  and  $H$  from above and below respectively, and to be distribution functions.

*Definition :* Let  $S$  be the unit square bounded below and to the left by the coordinate axes; then the line segment connecting  $(0, 0)$  with  $(1, 1)$  is its principal diagonal, the one connecting  $(0, 1)$  with  $(1, 0)$  its secondary diagonal. Let

$$(t) = \begin{cases} 0 & \text{if } t \leq 0, \\ t & \text{if } 0 \leq t \leq 1, \\ 1 & \text{if } t \geq 1. \end{cases}$$

Let

$$R_0(v, w) = (v)(w),$$

$$R_+(v, w) = \min\{(v), (w)\},$$

$$R_-(v, w) = \max\{0, (v)+(w)-1\}.$$

Let  $D$  be a class of disjoint open intervals on the abscissa of  $S$  and disjoint open intervals on the ordinate of  $S$ . For any point  $t$  on either axis, let  $t^D$  equal the lower limit of that interval (if any) of  $D$  to which  $t$  belongs, and equal to  $t$  otherwise. Finally, let

$$R_0^D(v, w) = (v^D)(w^D),$$

$$R_+^D(v, w) = \min\{(v^D), (w^D)\},$$

$$R_-^D(v, w) = \max\{0, (v^D)+(w^D)-1\}.$$

Evidently  $R_0$  is the uniform distribution over  $S$ , and  $R_+$  and  $R_-$  are the uniformly highest and lowest valued bivariate distribution functions over  $S$  whose marginals are uniform. We have, moreover,

*Lemma 2.1:*  $R_+[R_-]$  is that distribution function for which all the probability mass is concentrated along the principal [secondary] diagonal of the unit square  $S$  and is spread uniformly along that diagonal.  $R_+^D[R_-^D]$  is that distribution function for which all the probability mass is concentrated along the principal [secondary] diagonal of the unit square  $S$  and is spread uniformly along that diagonal, except that sections of the diagonal whose projections are contained in  $D$  have all the mass accumulated at that end point which is projected to upper limit of the relevant interval of  $D$ .

For  $Y$  and  $Z$  having continuous distributions  $G$  and  $H$ , it is well known that

$$P\{G(Y) \leq t\} = P\{H(Z) \leq t\} = (t).$$

If  $G$  or  $H$  have discontinuities, the ranges of  $G$  or  $H$  are not unit intervals, but exclude a class  $D$  of intervals and

$$P\{G(Y) \leq t\} = P\{H(Z) \leq t\} = (t^D).$$



# POSITIVE AND NEGATIVE DEPENDENCE OF TWO RANDOM VARIABLES

It follows that, if  $X_0$  has the distribution  $F_0$ ,  $X_+$  the distribution  $F_+$ , and  $X_-$  the distribution  $F_-$  then the couples

$$U_0 = (V_0, W_0) = (G(Y_0), H(Z_0)),$$

$$U_+ = (V_+, W_+) = (G(Y_+), H(Z_+)),$$

$$U_- = (V_-, W_-) = (G(Y_-), H(Z_-)),$$

have the distributions  $R_0^D$ ,  $R_+^D$  and  $R_-^D$ . In general, for  $X = (Y, Z)$  with a joint distribution  $F$ , we shall designate the distribution of the couple  $U = (V, W) = (G(Y), H(Z))$  by  $R$ .

The following theorem identifies  $F_+[F_-]$  with the completely positively [negatively] dependent distributions with marginals  $G$  and  $H$ , and in conjunction with the results of the next section, the continuous with the strictly monotone case. We shall say that  $r$  constitutes a *nondecreasing* [nonincreasing] relation between the points of two intervals if, for either interval,  $r$  associates with any point of one interval at least one point of the other, none of which smaller [larger] than any points  $r$  associates with a smaller point of the former interval;  $r$  is increasing [decreasing] if it is also one-to-one.

**Theorem 2.1:** *All the mass of the  $F_+[F_-]$  distribution is concentrated along a nondecreasing [nonincreasing] (possibly discontinuous) curve—that is, there is an almost sure nondecreasing relation between  $Y_+$  and  $Z_+$  [nonincreasing relation between  $Y_-$  and  $Z_-$ ]. Conversely, if between two random variables with marginal distributions  $G$  and  $H$  there is an almost sure nondecreasing [nonincreasing] relation, their joint distribution is  $F_+[F_-]$ . If  $F_0$  is continuous, the curve is strictly<sup>1</sup> monotone and the relation one-to-one, and conversely.*

*Proof:* Let  $S_0$  be the union of all rectangles  $S'$  in the  $(y, z)$ -plane such that if  $(y, z)$  and  $(y', z') \in S'$  with  $y < y', z < z'$ ,

$$G(y) = G(y') \text{ or } H(z) = H(z').$$

Then the joint distribution of any random variables with marginals  $G$  and  $H$  assigns no probability mass to  $S_0$ , and outside of  $S_0$  the functions  $G$  and  $H$  are still distribution functions and have unique inverses wherever some inverse exists.

Now if  $X_+$  has the distribution  $F_+$ ,  $U_+$  has the distribution  $R_+^D$ . So by Lemma 3.1 the locus of the probability mass of the  $F_+$  distribution is a nondecreasing (strictly increasing and continuous outside  $S_0$  if  $F_0$  is continuous) image of the principal diagonal of  $S$ , having no mass in  $S_0$ .

---

<sup>1</sup>The beginning and end points of intervals for which the curve contains no probability mass may, however, have the same  $y$  or  $z$  coordinates.



Conversely, if between  $Y$  and  $Z$  there is an almost sure monotone nondecreasing interval-valued relation  $Z = f(Y)$ , we have outside a set of  $t$  of zero  $G$ -measure

$$G(t) = P\{Y \leq t\} = P\{Z \leq \bar{f}(t)\} = H[\bar{f}(t)],$$

where  $\bar{f}(t)$  is the upper limit of the values of  $f(t)$ . Therefore,

$$H(Z) = H[f(Y)] = G(Y)$$

almost surely,

$$\begin{aligned} P\{G(Y) \leq v, H(Z) \leq w\} &= P\{G(Y) \leq v, G(Y) \leq w\} = \min\{(v^D), (w^D)\} \\ &= P\{G(Y) \leq v^D, H(Z) \leq w^D\}, \end{aligned}$$

and outside of  $S_0$   $P\{Y \leq y, Z \leq z\} = \min\{G(y), H(z)\}$ .

By definition of  $S_0$ , this relation must be preserved in  $S_0$ . If  $f$  is one-to-one, we have outside  $S_0$

$$f = H^{-1}G, \quad f^{-1} = G^{-1}H$$

with  $H^{-1}$  and  $G^{-1}$  free of jumps. So  $G$  and  $H$  have no jumps outside  $S_0$ , implying that  $F_0$  is continuous.

The proof for  $F_-$  and nonincreasing relations is similar.

### 3. CONTINUITY PROPERTIES

In Theorem 2.1 we obtained sharper results for the case in which  $F_0$  is continuous than for the general case. Continuity is also important if one wants to construct nonparametric tests of independence. We state without proof:

Theorem 3.1 : If  $F_0$  is continuous,  $F_+$  and  $F_-$  are also continuous.

Theorem 3.2 : Let  $X = (Y, Z)$  have the continuous distribution  $F$  with marginals  $G$  and  $H$ . Then (a)  $G$  and  $H$  are continuous, and (b) the distribution of  $U = (G(Y), H(Z))$  is continuous.

### 4. SOME PARAMETERS AND THEIR ESTIMATES

For any continuous bivariate distribution function  $F$  with marginals  $G$  and  $H$ , with  $\mu$  and  $\nu$  satisfying  $G(\mu) = H(\nu) = 1/2$ , and with  $F_0(y, z) = G(y)H(z)$ , define

$$\tau_1 = 4\left\{\int \int F(y, z)dF(y, z) - \int \int F_0(y, z)dF_0(y, z)\right\},$$

$$\tau_2 = 12 \int \int \{G(y) - \int G(\bar{y})dG(\bar{y})\}\{H(z) - \int H(\bar{z})dH(\bar{z})\}dF(y, z),$$

$$\tau_3 = 4\{F(\mu, \nu) - G(\mu)H(\nu)\}.$$

If the denominator is finite (it is positive for continuous  $F$ ) we define  $\rho$  by

$$\int \int \{y - \int \bar{y} dG(\bar{y})\} \{z - \int \bar{z} dH(\bar{z})\} dF(y, z) / [\int \{y - \int \bar{y} dG(\bar{y})\}^2 dG(y) \int \{z - \int \bar{z} dH(\bar{z})\}^2 dH(z)]^{\frac{1}{2}}.$$

$\rho$  is the (ordinary) correlation,  $\tau_1$  the difference sign correlation,  $\tau_2$  the grade correlation,  $\tau_3$  the medial correlation. The "natural" unbiased estimates of  $\tau_1$  and  $\tau_2$  based on a sample of  $n$  independent pairs of observations are given in Hoeffding (1948), the "natural" consistent estimate of  $\tau_3$  in Blomqvist (1950). To save space they will not be reproduced here; we shall denote them by  $T_{in}$  ( $i = 1, 2, 3$ ), and the sample correlation coefficient by  $R_n$ . The expectation of Spearman's (sample) rank correlation coefficient

$$T_{0n} = \frac{(n-2)T_{2n} + 3T_{1n}}{n+1}$$

will be denoted by  $\tau_0$ .

Let us first prove the following lemma, which may be of more general interest. (Professor Hoeffding kindly pointed out that for  $F$  absolutely continuous these relations are essentially contained in his doctoral dissertation.)

**Lemma 4.1:** *Let  $F$  be a bivariate distribution function with marginal distribution functions  $G$  and  $H$ . Then*

- (i)  $\int \int F dG dH = \int \int GH dF$  if  $G$  and  $H$  are continuous,
- (ii)  $\int \int (y - \int y dG)(z - \int z dH) dF = \int \int (F - GH) dy dz$  if the left-hand side, or  $\int y^2 dG$  and  $\int z^2 dH$ , exist.

*Proof:* Let  $V$  and  $W$  be continuous distribution functions. Then

$$\begin{aligned} \int \int V(y) W(z) dF(y, z) &= \int W(z) d_z \left( \int V(y) d_y F(y, z) \right) \\ &= W(z) \int V(y) d_y F(y, z) \Big|_{z=-\infty}^{z=\infty} - \int \left( \int V(y) d_y F(y, z) \right) dW(z). \end{aligned}$$

Now

$$\int V(y) d_y F(y, z) = V(y) F(y, z) \Big|_{y=-\infty}^{y=\infty} - \int F(y, z) dV(y) = H(z) - \int F(y, z) dV(y),$$

so

$$\begin{aligned} \int \int V(y) W(z) dF(y, z) &= \int H(z) dW(z) - \int \int F(y, z) dV(y) dW(z). \\ &= 1 - \int G(y) dV(y) - \int H(z) dW(z) + \int \int F(y, z) dV(y) dW(z). \end{aligned}$$

For  $V = G$ ,  $W = H$ , this reduces to (i). To obtain (ii), let  $h$  and  $k$  be positive integers,

let

$$y_{hk} = \begin{cases} -h & \text{if } y \leq -h \\ y & \text{if } -h \leq y \leq k \\ k & \text{if } y \geq k \end{cases}$$



and define  $z_{hk}$  similarly. Let  $V_{hk}(y) = (y_{hk} + h)/(h+k)$ ,  $W_{hk}(z) = (z_{hk} + h)/(h+k)$ .  
Let

$$c_{hk} = (h+k)^2 \int \int V_{hk}(y) W_{hk}(z) dF(y, z),$$

$$c'_{hk} = (h+k)^2 \int \int \left( V_{hk}(y) - \frac{h}{h+k} \right) \left( W_{hk}(z) - \frac{h}{h+k} \right) dF(y, z),$$

$$d_{hk} = (h+k)^2 \int \int F(y, z) dV_{hk}(y) dW_{hk}(z) = \int_{-h}^k \int_{-h}^k F(y, z) dy dz,$$

$$e_{hk} = (h+k) \int V_{hk}(y) dG(y),$$

$$e'_{hk} = (h+k) \int \left( V_{hk}(y) - \frac{h}{h+k} \right) dG(y),$$

$$f_{hk} = (h+k) \int W_{hk}(z) dH(z),$$

$$f'_{hk} = (h+k) \int \left( W_{hk}(z) - \frac{h}{h+k} \right) dH(z),$$

$$g_{hk} = (h+k) \int G(y) dV_{hk}(y) = \int_{-h}^k G(y) dy,$$

$$h_{hk} = (h+k) \int H(z) dW_{hk}(z) = \int_{-h}^k H(z) dz.$$

The previously obtained result then becomes (writing  $j = h + k$ )  
 $c_{hk} = j^2 - jg_{hk} - jh_{hk} + d_{hk}$ . As  $e_{hk} = j - g_{hk}$  and  $f_{hk} = j - h_{hk}$ ,  $c_{hk} - e_{hk} f_{hk} = d_{hk} - g_{hk} h_{hk}$ .  
Moreover,  $c'_{hk} - e'_{hk} f'_{hk} = c_{hk} - e_{hk} f_{hk}$ , so  $c'_{hk} - e'_{hk} f'_{hk} = d_{hk} - g_{hk} h_{hk}$ . The limit of the left-hand side as  $h$  and  $k \rightarrow \infty$  is, by definition of the Lebesgue-Stieltjes integral,

$$J = \int \int yz dF(y, z) - \int y dG(y) \int z dH(z)$$

and it exists by the hypothesis of (ii). Therefore the limit of the right-hand side,

$$\int \int F(y, z) dy dz - \int G(y) dy \int H(z) dz,$$

exists and equals  $J$ .

# POSITIVE AND NEGATIVE DEPENDENCE OF TWO RANDOM VARIABLES

The following theorem yields classes within which the statistics  $T_{in}$  ( $i = 0, 1, 2$ ) and  $R_n$  yield consistent tests of independence :

Theorem 4.1 : Let  $F$  be continuous,

$$\int \int_{F-F_0 < 0} dF = 0 \text{ or } \int \int_{F-F_0 < 0} dF_0 = 0,$$

and  $F > F_0$  for some  $x$ .

Then  $\tau_i [F] > 0$  for  $i = 0, 1, 2$ ;

and if  $\int \int y^2 dF < \infty$  and  $\int \int z^2 dF < \infty$ ,  $\rho[F] > 0$ .

*Proof:* We first show that the two first conditions imply that  $F - F_0 \geq 0$  everywhere. Suppose there would exist  $y_0, z_0$  such that  $F(y_0, z_0) - F_0(y_0, z_0) < 0$ . Let  $y', z'$  be the smallest numbers exceeding or equal to  $y_0$  and  $z_0$  respectively such that  $(y', z')$  is a point of increase of  $F$ , then there would exist  $y'', z''$  exceeding  $y'$  and  $z'$  respectively such that  $F - F_0 < 0$  throughout  $y_0 \leq y < y'', z_0 \leq z < z''$ , and since this rectangle contains a point of increase of  $F$  and  $F - F_0$  is continuous, this would contradict the hypothesis that  $\int \int_{F-F_0 < 0} dF = 0$ . On the other hand, let  $\bar{y}, \bar{z}$  be the largest numbers not exceeding  $y_0, z_0$  respectively such that  $(\bar{y}, \bar{z})$  is a point of increase of  $F_0$ , then there would exist  $\bar{\bar{y}}, \bar{\bar{z}}$  less than  $\bar{y}, \bar{z}$  respectively such that  $F - F_0 < 0$  throughout  $\bar{\bar{y}} < \bar{y} \leq y_0, \bar{\bar{z}} < \bar{z} \leq z_0$ , and this would contradict the hypothesis that  $\int \int_{F-F_0 < 0} dF_0 = 0$ .

Now,

$$\begin{aligned} \tau_1 &= 4 \int \int (F - F_0) d(F + F_0) = 4 \int \int_{F-F_0 > 0} (F - F_0) dF + 4 \int \int (F - F_0) dF_0 \\ &\geq 4 \int \int (F - F_0) dF_0 = \frac{1}{3} \tau_2, \end{aligned}$$

and

$$\int \int (F - F_0) dF_0 = \int_0^1 \int_0^1 \{R(v, w) - vw\} dv dw$$

has an integrand which is positive at at least one point and continuous, so that the integral is positive. Similarly

$$\text{cov} \{Y, Z\} = \int \int (F - F_0) dy dz$$

is positive.

## 5. THE VALUES OF THE PARAMETERS FOR POSITIVELY AND NEGATIVELY $\kappa$ -DEPENDENT VARIABLES

*Definition:* For  $0 \leq \kappa \leq 1$ , let  $F_\kappa^+ = (1 - \kappa)F_0 + \kappa F_+$ ,  $F_\kappa^- = (1 - \kappa)F_0 + \kappa F_-$ .



Let  $K[F_0]$  be the class of distributions obtainable from  $F_0$  by all such mixtures. In the case of one-sided alternatives we can consider the class obtainable by mixtures of  $F_0$  with  $F_+$  or with  $F_-$ ; one may then, if desired, distinguish notationally  $K^+[F_0]$  and  $K^-[F_0]$ .

Theorem 5.1 : Under  $F^+[F_-]$  (assuming existence and positiveness of second moments about the means)

$$0 \leq \rho(Y, Z) \leq \kappa [-\kappa \leq \rho(Y, Z) \leq 0];$$

the equality is only reached in case of a linear relation.

Proof : If, for example,  $X$  has the distribution  $F_+$ ,

$$\begin{aligned} \text{cov}\{Y, Z\} &= EYZ - EYEZ = (1-\kappa) \int \int yz dF_0 + \kappa \int \int yz dF_+ - \int y dG \int z dH \\ &= \kappa (\int \int yz dF_+ - \int \int yz dF_0) = \kappa \text{cov}\{Y_+, Z_+\} \geq 0. \end{aligned}$$

The last conclusion of the theorem is well-known, (see Cramér (1946) p. 265).

We now proceed to compute the values of the  $\tau_i$  under  $F_+$  and  $F_-$ , assumed continuous. Using Lemma 2.1 we get

$$\int \int F_+ dF_+ = \int_0^1 \int_0^1 R_+ dR_+ = \int_0^1 v dv = \frac{1}{2},$$

since  $R_+(v, w) = v$  along the diagonal  $v = w$ ;

$$\int \int F_+ dF_0 = \int_0^1 \int_0^1 R_+ dR_0 = \int_0^1 \int_v^1 v dw dv + \int_0^1 \int_0^v w dw dv = \frac{1}{3};$$

$$\int \int F_0 dF_+ = \int_0^1 \int_0^1 R_0 dR_+ = \int_0^1 v^2 dv = \frac{1}{3},$$

since  $R_0(v, w) = v^2$  along the diagonal  $v = w$ ;

$$\int \int F_- dF_- = \int_0^1 \int_0^1 R_- dR_- = \int_0^1 0 dw' = 0,$$

(where  $w' = 1-w$ ) since  $R_-(v, w) = 0$  along the diagonal  $v = w'$ ;

$$\int \int F_- dF_0 = \int_0^1 \int_0^1 R_- dR_0 = \int_0^1 \int_{1-v}^1 (v+w-1) dw dv = \frac{1}{6},$$

$$\int \int F_0 dF_- = \int_0^1 \int_0^1 R_0 dR_- = \int_0^1 w'(1-w') dw' = \frac{1}{6},$$

(where  $w' = 1-w$ ) since  $R_0(v, w) = v(1-w')$  equals  $w'(1-w')$  along the diagonal  $v = w'$ .

# POSITIVE AND NEGATIVE DEPENDENCE OF TWO RANDOM VARIABLES

Consequently,

$$\iint F_k^+ dF_k^+ = (\kappa^2 + 2\kappa + 3)/12, \quad \iint F_k^- dF_k^- = (-\kappa^2 - 2\kappa + 3)/12,$$

$$\iint F_k^+ dF_0 = \iint F_0 dF_k^+ = (\kappa + 3)/12, \quad \iint F_k^- dF_0 = \iint F_0 dF_k^- = (-\kappa + 3)/12,$$

so, in obvious notation,

$$\tau_1^+(\kappa) = (\kappa^2 + 2\kappa)/3, \quad \tau_1^-(\kappa) = -(\kappa^2 + 2\kappa)/3,$$

$$\tau_2^+(\kappa) = \kappa, \quad \tau_2^-(\kappa) = -\kappa.$$

Let  $\mu$  be a median of  $G$ , then, since  $G$  is continuous,  $G(\mu) = 1/2$ , and, as  $P\{G(Y_0) \leq 1/2\} = 1/2$ ,  $1/2$  is the median of the distribution of  $G(Y_0)$ . Similarly, if  $\nu$  is a median of  $H$ ,  $H(\nu) = 1/2$ , and  $1/2$  is the median of the distribution of  $H(Z_0)$ . Therefore,

$$F_k^+(\mu, \nu) = R_k^+(\frac{1}{2}, \frac{1}{2}) = (1 - \kappa)\frac{1}{4} + \kappa \cdot \frac{1}{2},$$

$$F_k^-(\mu, \nu) = R_k^-(\frac{1}{2}, \frac{1}{2}) = (1 - \kappa)\frac{1}{4} + \kappa \cdot 0,$$

so

$$\tau_3^+(\kappa) = \kappa, \quad \tau_3^-(\kappa) = -\kappa.$$

Theorem 5.2: Under  $F^+[F^-]$ , assumed continuous,  $\tau_i [-\tau_i]$  equals  $(\kappa^2 + 2\kappa)/3$  for  $i = 1$ , and  $\kappa$  for  $i = 2$  or  $3$ . Consequently, for  $F$  continuous and  $i = 0, 1, 2, 3$ ,

$$\tau_i = 1 \text{ or } \tau_i = -1$$

if the variates have an almost sure increasing or decreasing relation. The converse holds for  $i = 0, 1$  and  $2$ , but not for  $i = 3$ .

To show the converse for  $i = 1$ , merely observe<sup>1</sup> that for  $(Y_1, Y_2)$  and  $(Z_1, Z_2)$  independently distributed with common continuous distribution  $F$

$$\tau_1[F] = 2\text{Pr.}\{(Y_1 - Y_2)(Z_1 - Z_2) > 0\} - 1.$$

To show the converse for  $i = 2$  note that

$$\tau_2[F] = 12 \int \int (F - F_0) dF_0 = 12 \int_0^1 \int_0^1 \{R(v, w) - vw\} dv dw$$

is maximized for  $R = R_+$ , that is, for  $F = F_+$ , by the definition of  $F_+$ ; and if  $R'$  also maximizes  $\int_0^1 \int_0^1 \{R(v, w) - vw\} dv dw$ ,  $R' = R_+$  almost everywhere, so everywhere (since continuous). The proof for  $\tau_2[F] = -1$  is similar. One sees easily that  $\tau_0[F] = \pm 1$  implies  $\tau_1[F] = \pm 1$  (and also  $\tau_2[F] = \pm 1$  if  $n > 2$ ).

<sup>1</sup> I owe this observation to Professor W. Hoeffding.



$\tau_3[F] = \pm 1$  does not imply an almost sure monotone relation between  $Y$  and  $Z$ . For it is easy to visualize an almost sure non-monotone relation between  $Y$  and  $Z$ , the graph of which lies entirely in the two positive quadrants defined by the medians of  $Y$  and  $Z$  so that  $\tau_3[F] = 1$ .

As a contrast to these results  $\rho[F] = \pm 1$  if and only if all the mass of the  $F$  distribution lies along a straight line (assuming existence of second moments).

To the above corollary corresponds the following obvious property of the statistics  $T_{in}$ :

Theorem 5.3 : Let  $F_0$  be continuous. For  $i = 0, 1, 2, 3$ ,  $T_{in}$  equals 1 under  $F_+$  and  $-1$  under  $F_-$  almost surely.

## 6. UNBIASEDNESS, BOUNDS ON THE POWER AND ASYMPTOTIC POWER AGAINST ALTERNATIVES IN $\kappa$

Theorem 6.1 : Let  $F_0$  be continuous,  $i = 0, 1, 2, 3$ ,  $a = (a', a'')$ ,

$$s_{in}(a) = \{(x_1, \dots, x_n) : t_{in}(x_1, \dots, x_n) \leq a' \text{ or } \geq a''\},$$

$$P_{kn}^+\{S_{in}(a)\} = \int \dots \int_{s_{in}(a)} \prod_{j=1}^k dF_+(x_j) \prod_{j=k+1}^n dF_0(x_j),$$

and let  $t'_{in}$  be the largest and  $t''_{in}$  the smallest number for which  $0 < \alpha'_{in} = P\{T_{in} \leq t'_{in} | F_0\}$  and  $0 < \alpha''_{in} = P\{T_{in} \geq t''_{in} | F_0\}$  do not exceed  $\alpha' > 0$  and  $\alpha'' > 0$  respectively, but converge to these numbers. Here  $\alpha'_{in} + \alpha''_{in} = \alpha_{in}$  and  $\alpha' + \alpha'' = \alpha < 1/2$ . Denote the power of the  $(\alpha', \alpha'')$ -level independence test  $\mathcal{Z}_i$  (based on  $T_{in}$ ) against  $F_\kappa^+$  by  $\beta_n(\mathcal{Z}_i | F_\kappa^+)$ ,  $\kappa > 0$ . Define similarly the test against  $F_\kappa^-$  and its power. Then

$$(a) \quad \beta_n(\mathcal{Z}_i | F_\kappa^+) = (1-\kappa)^n \alpha_{in} + \kappa^n + \sum_{k=1}^{n-1} \binom{n}{k} \kappa^k (1-\kappa)^{n-k} P_{kn}^+\{S_{in}(t_{in})\},$$

(with the same holding for superscripts  $-$ );

$$(b) \quad 1 - (1-\kappa)^n (1-\alpha_{in}) \geq \beta_n(\mathcal{Z}_i | F_\kappa) > \alpha_{in} + \kappa^n (1-\alpha_{in}) \text{ (with } F_\kappa = F_\kappa^+ \text{ or } F_\kappa^-),$$

which exceeds  $\alpha_{in}$  for all continuous  $F_0$  so that the  $\mathcal{Z}_i$  are unbiased against alternatives in  $K[GH]$  with  $G, H$  continuous;

(c) the first inequality above is sharp for all but the lowest  $n$ ;

(d) the derivative of  $\beta_n(\mathcal{Z}_i | F_\kappa)$  with respect to  $\kappa$  at  $\kappa = 0$  equals  $n\epsilon_i$ , where  $\epsilon_i > 0$  and  $F_\kappa = F_\kappa^+$  or  $F_\kappa^-$ ;

(e) the same results hold for the correlation test if and only if there exist numbers  $a > 0$  and  $b$  such that  $G(y) = H(ay+b)$  [under  $F_{\kappa}^-$ :  $1-G(y) = H(-ay+b)$ ] for almost all  $y$ . (In this case the critical values  $t'_{inF_0}$ ,  $t'_{inF_0}$  depend on the particular  $F_0$ );

(f) the  $T_{in}$  are unbiased estimates of the  $\tau_{in}$  for  $i = 0, 1, 2$ ; and, under  $F_0$ ,  $ET_{in} = 0$  for  $i = 0, 1, 2$ , and 3.

*Proof:* (a) The proof follows by expansion of

$$\int \dots \int_{s_{in}(t_{in})} \prod_{j=1}^n dF_{\kappa}^+(x_j),$$

noting that, by Theorem 5.3,  $P_{nn}^+ \{S_{in}(t_{in})\}^n = 1$ .

(b) and (c) follow from the fact that the  $P_{kn}^+ \{S_{in}(t_{in})\}$  are nondecreasing functions of  $k$ , and strictly increasing ones,

(b) for small  $k$  (including at least  $k = 1$ ), and  
(c) (for all but the lowest  $n$ ): for  $k$  near  $n$  (including at least  $k = n-1$ ).

(d) follows from the fact cited in the proof of (b),

(e) and (f) are immediate.

We now examine the asymptotic power against positive  $\kappa'$  near 0. Let  $\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^x \exp(-t^2/2) dt$ ,  $\Phi(\delta') = \alpha'$ ,  $1 - \Phi(\delta'') = \alpha''$ . From the results of Hoeffding (1948) and Blomqvist (1950) and using the linearity in  $\kappa$  of  $F_{\kappa}^+$  and  $F_{\kappa}^-$ , it follows easily in this case that  $T_{in}$  is asymptotically normal (nonsingular except at  $\kappa = 1$ ). For  $R_n$ , when the second moments are finite, we note that  $ER_n = \rho + O(1/n)$ ,  $E(R_n - \rho)^2 = \frac{k}{n} + O(n^{-3/2})$  and the asymptotic distribution of  $(R_n - \rho)(k/n)^{-1/2}$  is normal with zero mean and unit variance if the fourth moments are finite [Cramér, 1946 p. 359 and p. 366], where  $k$  is a certain expression approximately equal to 1 when  $\kappa$  is small. Moreover, the moments are polynomials in  $\kappa$ , and by Theorem 4.1  $\rho$  is positive under  $F_{\kappa}^+$  and negative under  $F_{\kappa}^-$ . This gives the following result, in which independence from  $F_0$  should be noted.

**Theorem 6.2:** Consider  $(\alpha', \alpha'')$ -level tests for independence based on  $T_{in}$  ( $i = 0, 1, 2, 3$ ), and (if  $G$  and  $H$  have finite fourth moments) on  $R_n$ ; and let  $F_0$  be continuous. For alternatives in  $K^+[F_0]$  and  $K^-[F_0]$  for which  $\kappa'$  is near 0, their asymptotic power is

$$1 - \Phi(\delta'' - \sqrt{n}\kappa') + \Phi(\delta' - \sqrt{n}\kappa'),$$

whatever be  $F_0$ .



REFERENCES

- BLOMQUIST, N. (1950) : On a measure of dependence between two random variables. *Ann. Math. Stat.*, 21, 593-600.
- CRAMÉR, H. (1946) : *Mathematical Methods of Statistics*, Princeton University Press, Princeton.
- FRÉCHET, M. (1951) : Sur les tableaux de corrélation dont les marges sont données. *Ann. de l'Université de Lyon*, Section A, 3e Série, Fasc. 14, Paris, 53-77.
- HOEFFDING, W. (1948) : A class of statistics with asymptotically normal distributions. *Ann. Math. Stat.*, 19, 293-325.
- KONIJN, H. S. (1956) : On the power of certain tests for independence in bivariate populations. *Ann. Math. Stat.*, 27, 300-323. (Corrections in 29, 935-936).

*Paper received : April, 1957.*

# DEFINITION AND USE OF GENERALIZED PERCENTAGE POINTS

By JOHN E. WALSH

*Lockheed Aircraft Corporation, California\**

**SUMMARY.** Often what is assumed to be a random sample is in reality a set of independent observations each of which is from a separate statistical population, where some of these populations are noticeably different from the others. Also, situations where the observations are known to be from noticeably different populations frequently arise. Then "population percentage point" represents an undefined concept. This introduces the problem of generalizing the population percentage point concept to the situation of a set of independent observations from possibly different populations. The resulting generalized percentage point parameter should represent an intuitively understandable "average" of the percentage point properties of the populations involved. Also reasonably accurate point estimation, confidence intervals and significance tests should be available for this generalized percentage point parameter on the basis of the observations. This paper presents a generalized percentage point concept which satisfies these requirements for situations where the populations are continuous and do not differ greatly. Approximate methods are outlined for determining median estimates and confidence intervals for the values of generalized percentage points of the type presented.

## 1. INTRODUCTION

Statistical analysis based on a sample usually is much less difficult than a similar analysis based on independent observations from possibly different statistical populations. Consequently, there is a tendency to assume that a set of independent observations represents a sample, even when rough approximation to this situation is doubtful. Then a population percentage point investigation may not be meaningful since "population percentage point" is undefined if the observations are not from the same population. The purpose of this paper is to introduce a generalized percentage point concept which is applicable when the observations are from noticeably different populations. These generalized percentage points appear to be useful when the populations are different and reduce to the corresponding population percentage points when a single population is involved. Use of this generalized percentage point concept, combined with the application procedures developed in this paper, protects against the erroneous assumption of a random sample but incurs little penalty, if a random sample actually exists.

Let us consider some conditions which a generalized percentage point might be expected to satisfy when the data consist of a set of  $n$  independent observations from possibly different populations. These conditions are:

- (1) A generalized percentage point  $\theta_p$  is completely identified by its definition and the value of  $p$ .

---

\* The author is now with the System Development Corporation, Santa Monica, California, U.S.A.



- (2) The value of  $\theta_p$  depends on all of the  $n$  populations from which observations were drawn and represents some sort of continuous average of percentage point properties of these populations. This "average" should be directly related to the value of  $p$  and capable of intuitive interpretation.
- (3) If all the observations are from the same population,  $\theta_p$  equals the  $100p$  percent point for that population.
- (4) Reasonably accurate point estimation of  $\theta_p$  should be possible on the basis of the  $n$  observations.
- (5) Reasonably accurate confidence intervals and significance tests for  $\theta_p$  should be available on the basis of the  $n$  observations.

It is hardly to be expected that any formulated definition of  $\theta_p$  has the property that (1)-(5) are satisfied for all possible sets of  $n$  statistical populations. Conditions (4) and (5) limit the allowable situations. However, a  $\theta_p$  definition can be obtained with the property that conditions (1)-(5) hold for a rather general class of sets of  $n$  statistical populations.

Let us consider the restrictions on sets of  $n$  populations which are adopted in this paper. Stated in a qualitative manner, these restrictions are :

- (a) All the populations are continuous.
- (b) Let  $x_i$  be the observation from the  $i$ -th population, while  $p_i = Pr(x_i < \theta_p)$ , ( $i = 1, \dots, n$ ). The variation among the  $p_i$  is required not to be too large.

Restriction (a) is not completely necessary. That is, the  $\theta_p$  concept presented can satisfy all of (1)-(5) for situations, where one or more of the populations are not continuous. However, the analysis is greatly simplified if this restriction is adopted; moreover, many applied situations are such that (a) is acceptable. Condition (b) is stated in rather general terms. A technical specification of what is meant by the requirement that the variation among the  $p_i$  is not too large is given in the Definitions and Results section of this paper.

The quantity  $\theta_p$  presented in this paper is not necessarily unique. That is, there could be a range of values, all of which satisfy the specified requirements for  $\theta_p$ . This could happen, for example, if all the populations have probability density functions with ranges of zero values which are both preceded and followed by ranges of non-zero values. This lack of uniqueness property of  $\theta_p$  causes no difficulties in the analysis; moreover, lack of uniqueness can occur even if all the populations are the same.

The next section is titled Definitions and Results. This section contains the definition for  $\theta_p$  and a technical statement of restriction (b). A method for approximate median estimation of  $\theta_p$  and a procedure for obtaining approximate confidence



## DEFINITION AND USE OF GENERALIZED PERCENTAGE POINTS

intervals for  $\theta_p$  are also given. Significance tests can be obtained from these confidence intervals in the usual manner and are not considered explicitly. The final section is titled Verification and contains an outline of the basis for the results presented.

### 2. DEFINITIONS AND RESULTS

Let us consider a set of  $n$  independent observations whose values are denoted by  $x_1, \dots, x_n$ . Each observed value can occur from a possibly different statistical population, where these populations are unknown but fixed (i.e., not a sample from a universe of populations). The parameter  $\theta_p$  is defined as follows.

Definition of  $\theta_p$ : *The quantity  $\theta_p$  is the value, or set of values, which satisfies the relation*

$$\frac{1}{n} \sum_{i=1}^n p_i = \frac{1}{n} \sum_{i=1}^n \Pr(x_i < \theta_p) = p.$$

The quantity  $\theta_p$  has an uncomplicated intuitive interpretation. Namely,  $\theta_p$  is the 100 $p$ % point of the statistical population whose probability distribution is the arithmetic average of the distributions for the  $n$  observations considered. It is easily verified that the  $\theta_p$  concept defined here satisfies conditions (1)-(3).

The point estimate and confidence interval results presented in this paper are based on the assumption that each  $x_i$  is from a continuous population and that the variation among the values of the  $p_i$  is not too large. The allowable variation restriction is used in deriving confidence intervals for  $\theta_p$  and depends on the particular confidence interval considered. Let  $y(1), \dots, y(n)$  be the values of the  $x_i$  arranged according to increasing algebraic value. Thus  $y(1)$  is the smallest of the  $x_i$ , while  $y(n)$  is the largest. Then a confidence interval of the type considered is based on an arbitrary pair of the quantities  $y(0), y(1), \dots, y(n), y(n+1)$ ; here  $y(0) = -\infty$  and  $y(n+1) = \infty$ .

Suppose that  $y(n_1)$  and  $y(n_2+1)$  are the two quantities used, where  $0 \leq n_1 \leq n_2 \leq n$  and that the combination  $y(0), y(n+1)$  is excluded. Using  $q = 1 - p$ , let  $C_v(n_1, n_2)$  equal

$$\sum_{j=1}^2 \sum_{k=\max(1, v+n_j+j-n-1)}^{\min(v, n_j+j-1)} (-1)^{k+(v-1)j} \binom{n-v}{n_j+j-k-1} p^{n_j+j-k-1} q^{n-v-n_j-j+k+1}$$

$$\text{and } \sigma^2 = \frac{1}{n} \sum_{i=1}^n \left( p_i - \frac{1}{n} \sum_{i=1}^n p_i \right)^2 = \frac{1}{n} \sum_{i=1}^n (p_i - p)^2,$$

while  $\sigma_U$  is a known upper bound for the value of  $\sigma$ . Then restriction (b), the requirement that the variation among the  $p_i$  is not too large, can be expressed as follows.



Technical Statement of (b) : The value of  $\sigma_U$  is such that

$$n\sigma_U^2 | C_2(n_1, n_2) | \leq \min \left[ \sum_{r=n_1}^{n_2} \binom{n}{r} p^r q^{n-r}, 1 - \sum_{r=n_1}^{n_2} \binom{n}{r} p^r q^{n-r} \right]$$

and either or both of

$$\frac{1}{8} n\sigma_U^2 | C_4(n_1, n_2) | + \frac{1}{3} \sigma_U | C_3(n_1, n_2) | \leq \frac{1}{20} | C_2(n_1, n_2) |,$$

$$\frac{1}{8} n^2\sigma_U^4 | C_4(n_1, n_2) | + \frac{1}{3} n\sigma_U^3 | C_3(n_1, n_2) | + \frac{1}{2} n\sigma_U^2 | C_2(n_1, n_2) |$$

$$\leq \frac{1}{20} \min \left[ \sum_{r=n_1}^{n_2} \binom{n}{r} p^r q^{n-r}, 1 - \sum_{r=n_1}^{n_2} \binom{n}{r} p^r q^{n-r} \right]$$

are satisfied.

Next let us consider the statement and properties of the approximate confidence intervals for  $\theta_p$ . If the populations are all continuous and technical restriction (b) is satisfied, then for a very general class of situations

$$y(n_1) < \theta_p < y(n_2+1)$$

is a confidence interval for  $\theta_p$  with a confidence coefficient approximately equal to

$$\sum_{r=n_1}^{n_2} \binom{n}{r} p^r q^{n-r} + \frac{n}{4} (\sigma_U^2 + \sigma_L^2) C_2(n_1, n_2)$$

and a maximum confidence coefficient error of about

$$\frac{n}{4} (\sigma_U^2 - \sigma_L^2) | C_2(n_1, n_2) |.$$

Here  $\sigma_L$  is a known lower bound value for  $\sigma$  and is taken as zero if no better value is available.

*Confidence Interval Example.* As an illustration of the method, let  $n = 14$ ,  $p = .3$ ,  $\sigma_U = .05$ ,  $\sigma_L = 0$ ,  $n_1 = 2$ ,  $n_2 = 7$ . Then

$$\sum_{r=2}^7 \binom{14}{r} (.3)^r (.7)^{14-r} = .9210, \quad | C_2(2, 7) | = .1076,$$

$$| C_3(2, 7) | = .1662, \quad | C_4(2, 7) | = .2179.$$

# DEFINITION AND USE OF GENERALIZED PERCENTAGE POINTS

Then technical restriction (b) is satisfied since

$$.0038 \doteq n\sigma_U^2 |C_2(n_1, n_2)| \leq .0790 \doteq \min \left[ \sum_{n_1}^{n_2} \binom{n}{r} p^r q^{n-r}, 1 - \sum_{n_1}^{n_2} \binom{n}{r} p^r q^{n-r} \right],$$

$$.00372 \doteq \frac{1}{8} n\sigma_U^2 |C_4(n_1, n_2)| + \frac{1}{3} \sigma_U |C_3(n_1, n_2)| \leq .00538 \doteq \frac{1}{20} |C_2(n_1, n_2)|.$$

Thus, if all the populations are continuous,

$$y(2) < \theta_{.3} < y(8)$$

is a confidence interval for  $\theta_{.3}$  with a confidence coefficient of approximately .922 and a maximum confidence coefficient error of about .001.

Finally let us consider the procedure for obtaining a median estimate of  $\theta_p$ . This is based on the confidence interval results already presented. Thus it is required that all the populations are continuous and that technical restriction (b) is satisfied. Let  $n_1 = 0$  and determine the value of  $n_2$  which satisfies both

$$\sum_{r=0}^{n_2} \binom{n}{r} p^r q^{n-r} + \frac{n}{4} (\sigma_U^2 + \sigma_L^2) C_2(0, n_2) = \frac{1}{2} - \epsilon_1,$$

$$\sum_{r=0}^{n_2+1} \binom{n}{r} p^r q^{n-r} + \frac{n}{4} (\sigma_U^2 + \sigma_L^2) C_2(0, n_2+1) = \frac{1}{2} + \epsilon_2,$$

where  $\epsilon_1, \epsilon_2 \geq 0$ . Then

$$\frac{\epsilon_2}{\epsilon_1 + \epsilon_2} y(n_2+1) + \frac{\epsilon_1}{\epsilon_1 + \epsilon_2} y(n_2+2)$$

is an approximate median estimate for  $\theta_p$ . Here technical restriction (b) must be satisfied for both  $n_2$  and  $n_2+1$ . The value for  $n_2$  will often be in the vicinity of  $(n-1)p$  and this represents a first approximation to the value of  $n_2$ . The magnitudes of  $C_2(0, n_2)$ ,  $C_3(0, n_2)$ , and  $C_4(0, n_2)$  are all small for values of  $n_2$  near  $(n-1)p$ . In fact, the value of  $C_2(0, n_2)$  is zero if the integer  $n_2$  is equal to  $(n-1)p$ . Consequently, technical restriction (b) is nearly always satisfied in the case of median estimation of  $\theta_p$ .



*Median Estimate Example.* To illustrate this estimation method, let  $n = 20$ ,  $p = .2$ ,  $\sigma_U = .1$ ,  $\sigma_L = 0$ , and  $n_1 = 0$ . Then it is found that

$$\sum_{r=0}^3 \binom{20}{r} (.8)^r (.2)^{20-r} + .05 C_2(0, 3) = .4086,$$

$$\sum_{r=0}^4 \binom{20}{r} (.8)^r (.2)^{20-r} + .05 C_2(0, 4) = .6306.$$

Thus the value determined for  $n_2$  is 3 while  $\epsilon_1 = .0914$  and  $\epsilon_2 = .1306$ . Technical restriction (b) is easily shown to be satisfied for both  $n_2$  and  $n_2 + 1$ . Hence

$$.59y(4) + .41y(5)$$

is an approximate median estimate for  $\theta_{.2}$ .

This median method of estimation combined with the confidence interval results shows that the generalized percentage point concept presented here satisfies all the conditions (1)-(5) when restrictions (a) and (b) hold. Significance tests can easily be obtained on the basis of the confidence interval results. That is, let

$$y(n_1) < \theta_p < y(n_2 + 1)$$

be a confidence interval for  $\theta_p$  with confidence coefficient  $\alpha$ . Then the following rule is a significance test for comparing the specified value  $\theta_p^{(0)}$  with  $\theta_p$  and the significance level of this test is  $1 - \alpha$ .

Rule: *Reject that  $\theta_p^{(0)}$  which is equal to (or contained in)  $\theta_p$ , if either  $\theta_p^{(0)} \leq y(n_1)$  or  $\theta_p^{(0)} \geq y(n_2 + 1)$ .*

### 3. VERIFICATION

First let us consider the background for technical restriction (b). The combination of conditions

$$n\sigma_U^2 |C_2(n_1, n_2)| \leq \min \left[ \sum_{n_1}^{n_2} \binom{n}{r} p^r q^{n-r}, 1 - \sum_{n_1}^{n_2} \binom{n}{r} p^r q^{n-r} \right],$$

$$\frac{1}{8} n\sigma_U^2 |C_4(n_1, n_2)| + \frac{1}{3} \sigma_U |C_3(n_1, n_2)| \leq \frac{1}{20} |C_2(n_1, n_2)|$$

was taken directly from reference [ 1 ]. Consider  $n$  independent binomial observations with "success" probabilities of  $p_1, \dots, p_n$  and let  $X$  be the observed number of

# DEFINITION AND USE OF GENERALIZED PERCENTAGE POINTS

"successes." Then, on the basis of reference [ 1 ], if this combination of conditions holds,

$$Pr(n_1 \leq X \leq n_2) \doteq \sum_{r=n_1}^{n_2} \binom{n}{r} p^r q^{n-r} + \frac{n}{4} (\sigma_U^2 + \sigma_L^2) C_2(n_1, n_2)$$

with a maximum error of about

$$\frac{n}{4} (\sigma_U^2 - \sigma_L^2) | C_2(n_1, n_2) |.$$

Examination of the general expansion for  $Pr(x_1 \leq X \leq x_2)$  presented in Walsh (1955) strongly suggests that this approximate probability relation also holds if these conditions are replaced by the requirement

$$\frac{1}{8} n^2 \sigma_U^2 | C_4(n_1, n_2) | + \frac{1}{3} n \sigma_U^3 | C_3(n_1, n_2) | + \frac{1}{2} n \sigma_U^2 | C_2(n_1, n_2) |$$

$$\leq \frac{1}{20} \min \left[ \sum_{n_1}^{n_2} \binom{n}{r} p^r q^{n-r}, 1 - \sum_{n_1}^{n_2} \binom{n}{r} p^r q^{n-r} \right].$$

This condition asserts that the sum of the second, third and fourth terms of the  $Pr(n_1 \leq X \leq n_2)$  expansion amounts to at most 5 percent of both the first term and the probability complement of the first term. On the basis of the considerations of the probability complement of the first term. On the basis of the considerations of the reference [ 1 ], this strongly indicates that the third and higher order terms of the expansion can be neglected. To obtain mathematical rigor, the allowable populations are limited to those for which the approximate probability properties stated for  $Pr(n_1 \leq X \leq n_2)$  are satisfied. However, the important practical consideration is that these properties appear to be satisfied for virtually all situations of interest.

Next let us consider the confidence interval implications of technical restriction (b). Since the statistical populations are continuous and the observations are independent,

$$Pr[y(n_1) < \theta_p < y(n_2+1)] = Pr(n_1 \leq X \leq n_2)$$

and the stated confidence interval properties are verified.



Finally let us consider proof that  $C_2(0, n_2) = 0$  when  $n_2 = (n-1)p$ . By definition,

$$C_2(0, n_2) = \binom{n-2}{n_2-1} p^{n_2-1} q^{n-n_2-1} - \binom{n-2}{n_2} p^{n_2} q^{n-n_2-2}$$

$$= \frac{(n-2)! p^{n_2-1} q^{n-n_2-1}}{n_2! (n-n_2-1)!} \left[ n_2 - (n-1)p \right],$$

justifying the stated relation. Similar analysis shows that  $C_3(0, n_2)$  and  $C_4(0, n_2)$  tend to have small magnitudes when  $n_2 = (n-1)p$ .

#### REFERENCE

WALSH, J. E. (1955): Approximate probability values for observed number of 'successes' from statistically independent binomial events with unequal probabilities. *Sankhyā*, **15**, 281-290.

ERRATA FOR ABOVE REFERENCE : Expression for  $C_v(x_1, x_2)$  on p. 282: replace  $p_{n-v}^{(j-k-1+x_j)}$  by  $p_{n-v}^{(j-k-1+x_j)}$ . Line 3, p. 283: replace  $\frac{1}{2} \sigma^2 C_2(x_1, x_2)$  by  $\frac{n}{2} \sigma^2 C_2(x_1, x_2)$ . Last five lines of text of paper (p. 290): replace  $\sigma^2$  by  $\sigma^4$ .

*Paper received : February, 1957.*

# JOINT ASYMPTOTIC DISTRIBUTION OF U-STATISTICS AND ORDER STATISTICS

J. SETHURAMAN

*Indian Statistical Institute, Calcutta*  
and

B. V. SUKHATME

*Indian Council of Agricultural Research*

**SUMMARY.** It is shown under some mild restrictions that the joint distribution of a  $U$ -statistic (Hoeffding) and the  $a_n$ -th order statistic tends to (i) the bivariate normal distribution if  $\frac{a_n}{n} \rightarrow p, 0 < p < 1$ , (ii) the joint distribution of two independent variables, one of which is gamma and the other normal, in case  $a_n \rightarrow \text{constant}$  or  $n - a_n \rightarrow \text{constant}$ , (iii) the joint distribution of two independent normal variables if  $a_n \rightarrow \infty$  such that  $\frac{a_n}{n} \rightarrow 0$  or  $\frac{n - a_n}{n} \rightarrow 0$ . The above results are generalised to the case of several order statistics and several  $U$ -statistics. The generalisation to the case of several populations and generalised (Lehmann)  $U$ -statistics is also pointed out.

## 1. INTRODUCTION

Let  $x_1, \dots, x_n$  be  $n$  independent observations on a random variable  $X$ . Writing down the observations in increasing order we get  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . If  $\{a_n\}$  is a sequence of integers satisfying  $1 \leq a_n \leq n$ , then by the  $a_n$ -th order statistic we mean  $x_{(a_n)}$ . The random variable corresponding to it is  $X_{(a_n)}$ . We are interested in the joint asymptotic distribution of  $X_{(a_n)}$  and any  $U$ -statistic. Sukhatme (1957) has shown that the joint asymptotic distribution of  $X_{(\lfloor \frac{n}{2} \rfloor)}$  and any  $U$ -statistic with a bounded kernel, from a distribution whose density function is continuous at the median, is bivariate normal. We now proceed to prove the results stated in the summary.

## 2. THE CASE $a_n \rightarrow \infty; \frac{a_n}{n} \rightarrow p, 0 < p < 1$

**Theorem 1 :** Let  $x_1, \dots, x_n$  be  $n$  independent observations on a random variable  $X$  with distribution function  $F(x)$  and density function  $f(x)$ . Let  $f(x)$  be continuous at  $\theta$ , the  $p$ -th quantile of the population. Let  $y$  be the  $a_n$ -th order statistic from the sample where  $\frac{a_n}{n} \rightarrow p, 0 < p < 1$ . Let  $Y$  be the random variable corresponding to  $y$ . Let  $U_n$  be generated as a  $U$ -statistic from the bounded kernel

$$\psi(w_1, \dots, w_t). \quad \dots (2.1)$$



Let  $E(\psi(X_1, \dots, X_t)) = m.$

Then the joint distribution of

$$\{\xi = \sqrt{n}(Y - \theta), \eta = \sqrt{n}(U_n - m)\}$$

tends to the bivariate normal.

$$\text{Proof: Let } \varphi(x_1) = E(\psi(x_1, X_2, \dots, X_t)) \quad \dots (2.2)$$

$$E(\varphi(X) - m)^2 = \sigma^2 \quad \dots (2.3)$$

$$\left. \begin{aligned} \int_{-\infty}^{\theta} (\varphi(w) - m) f(w) dw &= m' \\ \int_{\theta}^{\infty} (\varphi(w) - m) f(w) dw &= m'' \\ \int_{-\infty}^{\theta} (\varphi(w) - m)^2 f(w) dw &= \sigma_1^2 \\ \int_{\theta}^{\infty} (\varphi(w) - m)^2 f(w) dw &= \sigma_2^2 \\ q &= 1 - p \end{aligned} \right\} \dots (2.4)$$

$$\text{Then we have } m' + m'' = 0, \quad \sigma_1^2 + \sigma_2^2 = \sigma^2. \quad \dots (2.5)$$

The characteristic function  $\varphi_n(t_1, t_2)$  of  $\left\{ \xi, \eta' = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\varphi(X_i) - m) \right\}$

$$\text{is given by } \varphi_n(t_1, t_2) = E\{\exp(it_1 \xi + it_2 \eta')\} = E_{\xi} [E\{\exp(it_1 \xi + it_2 \eta') \mid \xi\}]. \dots (2.6)$$

Then proceeding as in Sukhatme (1957) it turns out that

$$\begin{aligned} \varphi_n(t_1, t_2) &= \frac{n!}{(a_n - 1)! (n - a_n)!} \int_{-\infty}^{\infty} \exp\left(it_1 \xi + it_2 \frac{\varphi(y) - m}{\sqrt{n}}\right) f(y) dy \times \\ &\times \left[ \int_{-\infty}^y \exp\left(it_2 \frac{\varphi(w) - m}{\sqrt{n}}\right) f(w) dw \right]^{a_n - 1} \times \left[ \int_y^{\infty} \exp\left(it_2 \frac{\varphi(w) - m}{\sqrt{n}}\right) f(w) dw \right]^{n - a_n}. \end{aligned} \quad \dots (2.7)$$

Putting  $w = \theta + \frac{u}{\sqrt{n}}$  we find as in Sukhatme (1957) that

$$\left. \begin{aligned} \int_{-\infty}^y \exp\left(it_2 \frac{\varphi(w) - m}{\sqrt{n}}\right) f(w) dw &= p + it_2 \frac{m'}{\sqrt{n}} - \frac{t_2^2}{2n} \sigma_1^2 + \frac{\lambda}{\sqrt{n}} + \frac{\mu it_2}{n} + o\left(\frac{1}{n}\right) \\ \int_y^{\infty} \exp\left(it_2 \frac{\varphi(w) - m}{\sqrt{n}}\right) f(w) dw &= q + it_2 \frac{m''}{\sqrt{n}} - \frac{t_2^2}{2n} \sigma_2^2 - \frac{\lambda}{\sqrt{n}} - \frac{\mu it_2}{n} + o\left(\frac{1}{n}\right) \end{aligned} \right\} \dots (2.8)$$

$$\left. \begin{aligned} \text{where } \lambda &= \int_0^{\xi} f\left(\theta + \frac{u}{\sqrt{n}}\right) du \\ \mu &= \int_0^{\xi} \left\{ \varphi\left(\theta + \frac{u}{\sqrt{n}}\right) - m \right\} f\left(\theta + \frac{u}{\sqrt{n}}\right) du \end{aligned} \right\} \dots \quad (2.9)$$

We also note that for a fixed  $\xi$ ,  $\lambda \rightarrow \xi f(\theta)$  as  $n \rightarrow \infty$  since  $f(w)$  is continuous at  $\theta$ . Using (2.8) and simplifying (2.7) after expansion

$$\begin{aligned} & \log \left\{ \left[ \int_{-\infty}^y \exp\left(it_2 \frac{\varphi(w) - m}{\sqrt{n}}\right) f(w) dw \right]^{a_n - 1} \times \left[ \int_y^{\infty} \exp\left(it_2 \frac{\varphi(w) - m}{\sqrt{n}}\right) f(w) dw \right]^{n - a_n} \right\} \\ &= \text{const} - \frac{t_2^2}{2} \sigma^2 - \frac{\lambda^2}{2pq} + \frac{2it_2}{2} \lambda \left( \frac{m''}{q} - \frac{m'}{p} \right) + \frac{t_2^2}{2} \left( \frac{m'^2}{p} + \frac{m''^2}{q} \right) + o(1) \dots \quad (2.10) \end{aligned}$$

$$\begin{aligned} \text{Thus } \varphi_n(t_1, t_2) &= \text{const} \times \int_{-\infty}^{\infty} \exp \left[ it_1 \xi + it_2 \frac{\varphi\left(\theta + \frac{\xi}{\sqrt{n}}\right) - m}{\sqrt{n}} - \frac{t_2^2}{2} \sigma^2 + \right. \\ & \left. + \frac{t_2^2}{2} \left( \frac{m'^2}{p} + \frac{m''^2}{q} \right) + \frac{2it_2}{2} \lambda \left( \frac{m''}{q} - \frac{m'}{p} \right) - \frac{\lambda^2}{2pq} + o(1) \right] f\left(\theta + \frac{\xi}{n}\right) d\xi. \dots \quad (2.11) \end{aligned}$$

Now letting  $n \rightarrow \infty$ , and taking the lim sign inside the integral, which is valid in virtue of the bounded convergence theorem, we find that the right hand side of (2.11) without the constant tends to

$$\begin{aligned} & f(\theta) \int_{-\infty}^{\infty} \exp \left[ it_1 \xi - \frac{f^2(\theta) \xi^2}{2pq} + \frac{2it_2}{2} f(\theta) \xi \left( \frac{m''}{q} - \frac{m'}{p} \right) - \frac{t_2^2}{2} \sigma^2 + \frac{t_2^2}{2} \left( \frac{m'^2}{p} + \frac{m''^2}{q} \right) \right] d\xi \\ &= \sqrt{2\pi pq} \exp \left[ -\frac{t_2^2}{2} \sigma^2 - \frac{2t_1 t_2}{2} \left( \frac{m''}{q} - \frac{m'}{p} \right) \frac{pq}{f(\theta)} - \frac{t_1^2}{2} \frac{pq}{f^2(\theta)} \right] \end{aligned}$$

and the constant on the right hand side of (2.11), namely

$$\frac{1}{\int_{-\infty}^{\infty} \exp \left( -\frac{\lambda^2}{2pq} + o(1) \right) f\left(\theta + \frac{\xi}{\sqrt{n}}\right) d\xi} \rightarrow \frac{1}{\sqrt{2\pi pq}}.$$

$$\text{Thus } \varphi_n(t_1, t_2) \rightarrow \varphi(t_1, t_2) = \exp \left[ -\frac{t_2^2}{2} \sigma^2 - \frac{2t_1 t_2}{2} \left( \frac{m''}{q} - \frac{m'}{p} \right) \frac{pq}{f(\theta)} - \frac{t_1^2}{2} \frac{pq}{f^2(\theta)} \right]. \dots \quad (2.12)$$



Thus, we see that the asymptotic distribution of  $(\xi, \eta')$  is bivariate normal. From Hoeffding (1948) we see that  $E(\eta - t\eta')^2 \rightarrow 0$  as  $n \rightarrow \infty$ , so that  $\eta$  and  $t\eta'$  are asymptotically equivalent. Thus the asymptotic distribution of  $(\xi, \eta)$  is bivariate normal with zero means and asymptotic variances  $\frac{pq}{f^2(\theta)}$  and  $t^2\sigma^2$  the correlation coefficient

$$\text{being} \quad \left( \frac{m''}{q} - \frac{m'}{p} \right) \frac{\sqrt{pq}}{\sigma}. \quad \dots (2.13)$$

The above theorem was proved under the condition that  $\psi(w_1, \dots, w_l)$  is bounded. It is easy to show that the theorem holds good provided  $\psi(w_1, \dots, w_l)$  is bounded on any bounded interval of  $(w_1, \dots, w_l)$  and

$$E(\psi^3(X_1, \dots, X_l)) < \infty. \quad \dots (2.14)$$

This condition is sometimes more useful in practice.

The above theorems can be easily extended to the case of several  $U$ -statistics each of which satisfies one of the conditions (2.1) or (2.14). Again, the result can also be extended to the case of several order statistics. The last extension is quite straightforward, but the proof involves heavy algebra and hence will not be given here.

Another kind of extension of the above results is as follows :

Let  $(x_1, \dots, x_{n_1})$  and  $(y_1, \dots, y_{n_2})$  be independent observations on two independent random variables  $X$  and  $Y$  respectively. Let  $\{a_n\}, \{b_n\}$  be two sequences of integers and let  $Z_1, Z_2$  be the random variables corresponding to  $x_{(a_{n_1})}$  and  $y_{(b_{n_2})}$ . Let  $\frac{a_n}{n} \rightarrow p_1$ ,  $\frac{b_n}{n} \rightarrow p_2$   $0 < p_1, p_2 < 1$ . Let  $\theta_1$  and  $\theta_2$  be the  $p_1$ -th,  $p_2$ -th, quantiles of  $X$  and  $Y$  respectively.

Let  $U_{n_1, n_2}$  be a generalised (Lehmann)  $U$ -statistic with kernel  $\psi(x_1, \dots, x_{t_1}; y_1, \dots, y_{t_2})$  which is either bounded or bounded in any bounded interval of its arguments and possesses a third moment. If  $E(\psi) = m$  then the joint distribution of

$$\{\sqrt{n_1}(Z_1 - \theta_1), \sqrt{n_2}(Z_2 - \theta_2), \sqrt{n_1 n_2}(U_{n_1, n_2} - m)\}$$

tends to the trivariate normal distribution as  $n_1, n_2 \rightarrow \infty$  such that  $\frac{n_1}{n_2} \rightarrow c$ ,  $0 < c < \infty$ .

The proof depends on the fact that

$$\sqrt{n_1}(U_{n_1, n_2} - m) \text{ and } \frac{1}{\sqrt{n_1}} t_1 \sum_{i=1}^{n_1} \psi_1(X_i) + \frac{\sqrt{n_1}}{\sqrt{n_2}} \frac{t_2}{\sqrt{n_2}} \sum_{j=1}^{n_2} \psi_2(Y_j)$$

are asymptotically equivalent (see Fraser, 1957) where

$$\psi_1(x_1) = E(\psi(x_1, X_2, \dots, X_{t_1}; Y_1, \dots, Y_{t_2}) - m$$

and

$$\psi_2(y_1) = E(\psi(X_1, \dots, X_{t_1}; y_1, Y_2, \dots, Y_{t_2}) - m.$$

3. THE CASE  $a_n \rightarrow s$  OR  $n - a_n \rightarrow (r-1)$  WHERE  $s$  AND  $r$  ARE CONSTANTS

Since  $\{a_n\}$  is a sequence of integers, it is obvious that  $a_n \rightarrow s$  and  $n - a_n \rightarrow (r-1)$  implies that after a certain stage,  $a_n = s$  and  $n - a_n = (r-1)$  respectively, so that we may take them to be  $s$  and  $(r-1)$  respectively for all  $n$ . Thus we will have to find the joint asymptotic distribution of the  $s$ -th and  $(n-r+1)$ -th order statistics and  $U$ -statistic.

Theorem 2 : Let  $x_1, \dots, x_n$  be  $n$  independent observations made on a random variable  $X$  with a distribution function  $F(x)$  which is continuous. Let  $y$  and  $z$  be the  $s$ -th and  $(n-r+1)$ -th order statistics of the sample. Let  $U_n$  be a  $U$ -statistic generated out of a bounded symmetric kernel  $\psi(w_1, \dots, w_t)$ .

$$\text{Let} \quad \left. \begin{aligned} E(\psi(X_1, \dots, X_t)) &= m \\ \xi &= n F(Y) \\ \eta &= n(1-F(Z)) \\ \zeta &= \sqrt{n}(U_n - m) \end{aligned} \right\}. \quad \dots \quad (3.1)$$

Then, in the asymptotic distribution of  $(\xi, \eta, \zeta)$  the variables are independent and the marginal distributions are gamma, gamma and normal, respectively.

*Proof:* It is obvious that we need prove the theorem for the case of the rectangular distribution only.

Let

$$\varphi(x_1) = E(\psi(x_1, X_2, \dots, X_t)); \quad E(\varphi(X) - m)^2 = \sigma^2 \quad \dots \quad (3.2)$$

$$\zeta' = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\varphi(X_i) - m).$$

Then the characteristic function of  $(\xi, \eta, \zeta')$  is given by

$$\varphi_n(t_1, t_2, t_3) = E\{\exp(it_1\xi + it_2\eta + it_3\zeta')\} = E_{\xi, \eta}[E\{\exp(it_1\xi + it_2\eta + it_3\zeta') | \xi, \eta\}]. \quad \dots \quad (3.3)$$

Arguing as before

$$\begin{aligned} \varphi_n(t_1, t_2, t_3) &= \frac{n!}{(s-1)!(n-r-s)!(r-1)!} \int \int_{0 \leq \xi + \eta \leq n} \exp \left[ it_1\xi + it_2\eta + it_3 \left( \frac{\varphi\left(\frac{\xi}{n}\right) - m}{\sqrt{n}} + \frac{\varphi\left(1 - \frac{\eta}{n}\right) - m}{\sqrt{n}} \right) \right] \\ &\quad \times \left[ \int_0^{\xi/n} \exp \left( it_3 \frac{\varphi(w) - m}{\sqrt{n}} \right) dw \right]^{s-1} \times \left[ \int_{\xi/n}^{1-\eta/n} \exp \left( it_3 \frac{\varphi(w) - m}{\sqrt{n}} \right) dw \right]^{r-s} \times \\ &\quad \times \left[ \int_{1-\eta/n}^1 \exp \left( it_3 \frac{\varphi(w) - m}{\sqrt{n}} \right) dw \right]^{r-1} \frac{d\xi d\eta}{n^2}. \quad \dots \quad (3.4) \end{aligned}$$



It is easily seen that

$$\left. \begin{aligned} \int_0^{\xi/n} \exp\left(it_3 \frac{\varphi(w)-m}{\sqrt{n}}\right) dw &= \frac{\xi}{n} + o\left(\frac{1}{n}\right) \\ \int_{\xi/n}^{1-\eta/n} \exp\left(it_3 \frac{\varphi(w)-m}{\sqrt{n}}\right) dw &= 1 - \frac{t_3^2}{2n} \sigma^2 - \frac{\xi}{n} - \frac{\eta}{n} + o\left(\frac{1}{n}\right) \\ \int_{1-\eta/n}^1 \exp\left(it_3 \frac{\varphi(w)-m}{\sqrt{n}}\right) dw &= \frac{\eta}{n} + o\left(\frac{1}{n}\right) \end{aligned} \right\} \dots \quad (3.5)$$

so that

$$\begin{aligned} \varphi_n(t_1, t_2, t_3) &= \frac{n!}{(s-1)!(n-r-s)!(r-1)!n^{r+s}} \int \int_{0 \leq \xi + \eta \leq n} \exp \left[ it_1 \xi + it_2 \eta + it_3 \left( \frac{\varphi\left(\frac{\xi}{n}\right) - m}{\sqrt{n}} + \right. \right. \\ &\quad \left. \left. + \frac{\varphi\left(1 - \frac{\eta}{n}\right) - m}{\sqrt{n}} \right) \right] \times [\xi + o(1)]^{s-1} \times \left[ 1 - \frac{t_3^2}{2n} \sigma^2 - \frac{\xi}{n} - \frac{\eta}{n} + o\left(\frac{1}{n}\right) \right]^{n-r-s} \times [\eta + o(1)]^{r-1} d\xi d\eta. \end{aligned}$$

... (3.6)

Now letting  $n \rightarrow \infty$ , and taking the lim sign inside the integral which is valid in virtue of the bounded convergence theorem, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \varphi_n(t_1, t_2, t_3) &= \text{const} \times \int_0^\infty \int_0^\infty \exp \left[ it_1 \xi + it_2 \eta - \frac{t_3^2}{2} \sigma^2 - \xi - \eta \right] \xi^{s-1} \eta^{r-1} d\xi d\eta \\ &= \text{const} \times \frac{1}{(1-it_1)^s} \times \frac{1}{(1-it_2)^r} \times e^{-\frac{t_3^2}{2} \sigma^2} \end{aligned}$$

... (3.7)

where again, we can easily see that the constant is unity and hence the theorem is proved. Extensions to the case of several  $U$ -statistics and generalised  $U$ -statistics are obvious. Incidentally we note that the  $s$ -th order statistic and the  $(n-r+1)$ -th order statistic are independent in the limit, if  $r$  and  $s$  are constants.

#### 4. THE CASE $a_n \rightarrow \infty$ ; $\frac{a_n}{n} \rightarrow 0$ or $\frac{n-a_n}{n} \rightarrow 0$

**Theorem 3 :** Let  $x_1, \dots, x_n$  be  $n$  independent observations on a random variable  $X$  with distribution function  $F(x)$  which is continuous. Let  $\{a_n\}, \{b_n\}$  be two sequences of integers such that

$$1 \leq a_n < b_n \leq n \text{ and } a_n \rightarrow \infty, b_n \rightarrow \infty, \frac{a_n}{n} \rightarrow 0, \frac{b_n}{n} \rightarrow 1, \frac{c_n}{a_n} < K \text{ a constant, } \dots \quad (4.1)$$

where  $c_n = n - b_n$ .

# JOINT ASYMPTOTIC DISTRIBUTION OF U-STATISTICS AND ORDER STATISTICS

Let  $U_n$  be any  $U$ -statistic generated by a bounded kernel  $\psi(w_1, \dots, w_l)$  and let

$$\left. \begin{aligned} \varphi(x_1) &= E(\psi(x_1, X_2, \dots, X_l)) \\ E(\varphi(X)) &= m \\ E(\varphi(X) - m)^2 &= \sigma^2 \end{aligned} \right\} \quad \dots (4.2)$$

Then the asymptotic distribution of  $(\xi, \eta, \zeta)$  where

$$\left. \begin{aligned} \xi &= \frac{n}{\sqrt{a_n}} \left( F(X_{(a_n)}) - \frac{a_n}{n} \right) \\ \eta &= \frac{n}{\sqrt{c_n}} \left( \frac{b_n}{n} - F(X_{(b_n)}) \right) \\ \zeta &= \sqrt{n}(U_n - m) \end{aligned} \right\} \quad \dots (4.3)$$

is given by the density function, constant  $\times e^{-\frac{1}{2}(\xi^2 + \eta^2 + \frac{\zeta^2}{\sigma^2})}$ .

*Proof:* It is obvious that we need prove the theorem for the rectangular case only. If  $\zeta' = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\varphi(X_i) - m)$  then  $\varphi_n(t_1, t_2, t_3)$  the characteristic function of  $(\xi, \eta, \zeta')$  is given by

$$\begin{aligned} \varphi_n(t_1, t_2, t_3) &= E\{\exp(it_1\xi + it_2\eta + it_3\zeta')\} \\ &= E_{\xi, \eta}[E\{\exp(it_1\xi + it_2\eta + it_3\zeta') \mid \xi, \eta\}]. \end{aligned} \quad \dots (4.4)$$

Proceeding on the same lines as in the previous cases

$$\begin{aligned} \varphi_n(t_1, t_2, t_3) &= \frac{n!}{(a_n - 1)!(n - a_n - c_n + 1)!(c_n)!} \frac{\sqrt{a_n c_n}}{n^2} \times \\ &\times \int \int \exp \left[ it_1 \xi + it_2 \eta + it_3 \left( \frac{\varphi \left( \frac{a_n}{n} + \frac{\sqrt{a_n}}{n} \xi \right) - m}{\sqrt{n}} + \frac{\varphi \left( 1 - \frac{c_n}{n} - \frac{\sqrt{c_n}}{n} \eta \right) - m}{\sqrt{n}} \right) \right] \times \\ &0 \leq \sqrt{a_n} \xi + \sqrt{c_n} \eta \leq n - a_n - c_n \\ &\times \left[ \int_0^{\frac{a_n}{n} + \frac{\sqrt{a_n}}{n} \xi} \exp \left( it_3 \frac{\varphi(w) - m}{\sqrt{n}} \right) dw \right]^{a_n - 1} \times \\ &\times \left[ \int_{\frac{a_n}{n} + \frac{\sqrt{a_n}}{n} \xi}^{1 - \frac{c_n}{n} - \frac{\sqrt{c_n}}{n} \eta} \exp \left( it_3 \frac{\varphi(w) - m}{\sqrt{n}} \right) dw \right]^{n - a_n - c_n + 1} \times \\ &\times \left[ \int_{1 - \frac{c_n}{n} - \frac{\sqrt{c_n}}{n} \eta}^1 \exp \left( it_3 \frac{\varphi(w) - m}{\sqrt{n}} \right) dw \right]^{c_n} d\xi d\eta \end{aligned} \quad \dots (4.5)$$



which on simplifying, as done in previous paras, reduces to

$$\varphi_n(t_1, t_2, t_3) = \text{const} \times \int \int_{\sqrt{a_n}\xi + \sqrt{c_n}\eta \leq n - a_n - c_n} \exp \left[ it_1\xi + it_2\eta + it_3 \left( \frac{\varphi \left( \frac{a_n}{n} + \frac{\sqrt{a_n}}{n} \xi \right) - m}{\sqrt{n}} + \frac{\varphi \left( 1 - \frac{c_n}{n} - \frac{\sqrt{c_n}}{n} \eta \right) - m}{\sqrt{n}} - \frac{t_3^2}{2} \sigma^2 - \frac{\xi^2}{2} - \frac{\eta^2}{2} + o(1) \right] d\xi d\eta. \quad \dots (4.6)$$

Letting  $n \rightarrow \infty$ , (the limits for  $\xi, \eta$  become  $(0, \infty), (0, \infty)$  because  $\frac{c_n}{a_n} < K$ ), the above integral without the constant tends to

$$\int_0^\infty \int_0^\infty \exp \left[ -\frac{t_3^2}{2} \sigma^2 + it_1\xi + it_2\eta - \frac{\xi^2}{2} - \frac{\eta^2}{2} \right] d\xi d\eta = 2\pi \exp \left[ -\frac{t_1^2}{2} - \frac{t_2^2}{2} - \frac{t_3^2}{2} \sigma^2 \right] \quad \dots (4.7)$$

and the constant which is

$$\frac{1}{\int \int_{\sqrt{a_n}\xi + \sqrt{c_n}\eta \leq n - a_n - c_n} \exp(-\xi^2/2 - \eta^2/2 + o(1)) d\xi d\eta} \rightarrow \frac{1}{2\pi}.$$

Thus 
$$\varphi_n(t_1, t_2, t_3) \rightarrow \exp \left[ -\frac{t_1^2}{2} - \frac{t_2^2}{2} - \frac{t_3^2}{2} \sigma^2 \right] \quad \dots (4.8)$$

and hence the theorem is proved.

The extensions to several  $U$ -statistics and generalised  $U$ -statistics are immediate. We note that the  $a_n$ -th and  $b_n$ -th order statistics when suitably standardised are asymptotically normally and independently distributed if condition (4.1) holds.

In section 4 and section 5 we have shown that  $U_n$  and  $F(Y)$ , where  $Y$  is the  $a_n$ -th order statistic, have a certain asymptotic distribution. To make a similar statement about  $U_n$  and  $Y$  we need the following, (in case  $\frac{a_n}{n} \rightarrow 0$ ).

$F(x)$  has a density function  $f(x)$  and  $\lim_{F(x) \rightarrow 0} f(x)$  exists and is not equal to zero. For an example where this occurs, we may cite the exponential distribution.

5. REMARKS<sup>1</sup>

In the above proofs the use of characteristic functions has partly blurred out the picture of the exact process by which the limit distributions were attained. To gain some insight into this aspect we reason as follows: Suppose we fix the  $a_n$ -th order statistic, i.e., the normalised variable corresponding to it.

If  $\eta' = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\varphi(X_i) - m)$  we see that it splits up into 3 parts. One part being the sum of  $(a_n - 1)$  independent random variables where  $x$  ranges on  $(-\infty, x_{(a_n)})$ , another part being the sum of  $(n - a_n)$  independent random variable where  $x$  ranges on  $(x_{(a_n)}, \infty)$

and a third part fixed at  $\frac{\varphi(x_{(a_n)}) - m}{\sqrt{n}}$ . From an application of the Central Limit

Theorem we see that the limit of the conditional distribution of  $\eta'$  for a fixed  $\xi$  is normal. It can also be shown, after some algebra, that the mean and variance of this distribution are

$$\left( -f(\theta) \xi \left( \frac{m''}{q} - \frac{m'}{p} \right), \sigma^2 - \left( \frac{m'^2}{p} + \frac{m''^2}{q} \right) \right) \quad \text{in the case 2}$$

$$(0, \sigma^2) \quad \text{in the case 3}$$

$$(0, \sigma^2) \quad \text{in the case 4}$$

and in each of these cases we know the limiting marginal distribution of  $\xi$  so that the nature of the limiting joint distribution can be concluded to be bivariate normal in case 2, the distribution of independent normal and gamma variables in case 3, and the distribution of two independent normal variables in case 4. The conclusion can be justified using the following

*Lemma : Let  $(X_n, Y_n)$  be a sequence of random variables. Let  $F_n(y/x)$ , the conditional distribution of  $Y_n$  given  $X_n = x$ , tend weakly to a distribution  $F(y/x)$ . Also let  $G_n(x)$  the marginal distribution of  $X_n$ , tend weakly to a distribution function  $G(x)$ . Then under some conditions  $F_n(x, y)$  the joint distribution of  $(X_n, Y_n)$  tends to the distribution  $\int_{-\infty}^x F(y/x) dG(x)$ .*

A set of sufficient conditions are (1)  $F(y/x)$  is continuous in  $y$  for each  $x$ , (2)  $G_n(x)$ ,  $G(x)$  admit of probability densities  $g_n(x)$ ,  $g(x)$  respectively, and  $g_n(x) \rightarrow g(x)$  uniformly in any bounded interval of  $x$ . Both, that these conditions are sufficient and that these conditions are satisfied in our case, can be easily verified.

---

<sup>1</sup> We are grateful to Dr. S. K. Mitra of the Indian Statistical Institute for these remarks.



# 6. ACKNOWLEDGEMENTS

The authors wish to thank Dr. D. Basu of the Indian Statistical Institute for the several suggestions and helpful discussions during the course of the work presented here.

## REFERENCES

- FRASER, D. A. S. (1957): *Non-parametric Methods in Statistics*, 257. John Wiley & Sons, New York.
- HOEFFDING, W. (1948): A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.*, Series B, 19, 293-325.
- SUKHATME, B. V. (1957): Joint asymptotic distribution of the median and a *U*-statistic. *J. Roy. Stat. Soc.*, Series B, 19, 144-148.

*Paper received : June, 1958.*

*Revised : December, 1958.*

# SOME SAMPLING SYSTEMS PROVIDING UNBIASED RATIO ESTIMATORS

By N. S. NANJAMMA, M. N. MURTHY, and V. K. SETHI

*Indian Statistical Institute, Calcutta*

**SUMMARY.** In this paper, modifications of many of the selection procedures commonly adopted in practice, namely, equal probability sampling, varying probability sampling, stratified sampling and multi-stage sampling have been proposed, which, while retaining the form of the usual ratio estimators, make them unbiased. For many of the situations commonly met with in practice, this modification of a given sampling scheme consists essentially in first selecting one unit with probability proportional to its value of the characteristic occurring in the denominator of the ratio and then the remaining units in the sample according to the original scheme of sampling. The expressions for unbiased variance estimators of the unbiased ratio estimators have been given for some of the more important sampling schemes. Further the selection and estimation procedure which provide unbiased ratio estimators in the case of a certain general class of population parameters together with the expressions for its sampling variance and variance estimators have also been considered.

## 1. INTRODUCTION

As the relationship between two characteristics is usually of much interest, estimation of ratios of certain population parameters has become quite important in a large number of surveys. The method of ratio estimation is also being used to estimate population totals, since a ratio estimator is more efficient than the conventional unbiased estimator under certain circumstances not uncommon in actual practice. The usual procedure of using the ratio method in estimating any population ratio or total has been to take the ratio of unbiased estimators of the numerator and the denominator and in the latter case multiply it by the population total of the supplementary variate taken in the denominator. A disadvantage of this method is that the estimator so obtained is biased for many of the selection procedures commonly adopted in surveys. Further a completely satisfactory (at least to the present authors) treatment of the errors and biases of a ratio estimator is not yet available. For small samples, at least, the bias is not likely to be small.

In recent years attempts have been made to give selection and estimation procedures which provide unbiased ratio estimators. Lahiri (1951) has given a method of selecting a sample with probability proportional to its total size (pps) (sum of the sizes of the units in the sample) which is essentially similar to his method of selecting a unit with pps, namely, of selecting a unit with equal probability and including that unit in the sample if a number chosen at random from one to an upper bound of the units is less than or equal to the size of the selected unit. By 'size' here is meant the value of the supplementary variate under consideration. Obviously this method avoids the need for completely enumerating all possible samples and finding their total sizes and the cumulated sizes. Once a sample is chosen with pps it is easy to obtain an unbiased ratio estimator. The disadvantage of the selection procedure given by Lahiri is that it involves rejection of some draws.



Midzuno (1952) and Sen (1952) have independently given a simple procedure for obtaining a sample with pps. Their method consists in selecting one unit with pps and the rest with equal probability without replacement from the remaining units of the population. It may be observed that Lahiri's method of selecting one unit with pps could profitably be used in the selection procedure given by Midzuno and Sen.

In the case of stratified sampling Lahiri has pointed out that his method could be applied to select a sample with probability proportional to  $\sum_{s=1}^k N_s \bar{x}_s$  where  $k$  is the number of strata,  $N_s$  the number of units in the  $s$ -th stratum and  $\bar{x}_s$  the  $s$ -th stratum sample mean of the supplementary variate under consideration with a view to get an unbiased ratio estimator. Des Raj (1954) has given the expressions for the variance and an unbiased variance estimator of the ratio estimator in the case of a multi-stage design where the sample of first stage units is selected with pps.

So far the selection procedures providing unbiased ratio estimators have been given only for simple designs and that too for a very restricted class of parameters. In the next few sections, modifications of many of the selection procedures commonly adopted in practice, namely, equal probability sampling, varying probability sampling, stratified sampling and multi-stage sampling, have been given which, while retaining the form of the usual biased ratio estimators, make them unbiased. The expressions for the unbiased variance estimator for some of the more important cases of ratio estimators are also given. Further the selection and estimation procedures for obtaining unbiased ratio estimators in the case of a certain general class of population parameters together with the expressions for the sampling variance and the variance estimator are given in the last few sections.

For many of the situations commonly met with in practice, the modification of a given sampling scheme referred to above which provides unbiased ratio estimator consists essentially in first selecting one unit with probability proportional to its value of the variate occurring in the denominator of the ratio and then the remaining units in the sample according to the original scheme of sampling. For many of the sampling schemes considered in this paper it might be expected that in large samples the bias of the conventional ratio estimator is unlikely to be large, since the form of the ratio estimator is the same in the case of the biased and the unbiased ratio estimators and the sample based on the original sampling scheme and that on the modified scheme could be made the same but for a difference of one unit at the most.

In this paper the estimator and its variance estimator have been given in the case of estimating the ratio  $R = \frac{Y}{X}$  where  $Y$  and  $X$  are the population totals for two characters. The ratio estimator and its variance estimator for estimating  $Y$  can be obtained by multiplying the corresponding estimators in the case of estimation of  $R$  by  $X$  and  $X^2$  respectively.

## 2. EQUAL PROBABILITY SAMPLING

In the case of *unstratified unistage sampling with equal probability without replacement*, as has been mentioned earlier, Midzuno and Sen have suggested the method of selecting one unit with probability proportional to  $x(\text{ppx})$ , where  $x$  is the value of the variate occurring in the denominator of the ratio and the rest of the  $n-1$  units in the sample from the remaining  $N-1$  units in the population with equal probability without replacement. The probability of getting a particular sample  $s$  by this approach is given by

$$P(s) = \frac{1}{\binom{N}{n}} \frac{\bar{x}}{\bar{X}} \quad \dots (2.1)$$

where  $\bar{x}$  and  $\bar{X}$  are the sample and the population means respectively. Hence the estimator

$$\hat{R} = \frac{\bar{y}}{\bar{x}} \quad \dots (2.2)$$

where  $\bar{y}$  is the sample mean of the variate  $y$  is an unbiased estimator of the ratio  $R = Y/X$ . Though this estimator resembles the usual ratio estimator in the case of equal probability sampling without replacement, this is unbiased while the latter is not. An unbiased estimator of the variance of  $\hat{R}$  given in (2.2) is given by

$$\hat{V}(\hat{R}) = \hat{R}^2 - \frac{\sum_{i=1}^n y_i^2 + 2 \frac{N-1}{n-1} \sum_{i>j}^n y_i y_j}{Nn \bar{x} \bar{X}}. \quad \dots (2.3)$$

It can be seen that the efficiency of the unbiased estimator given in (2.2) will be greater than, equal to, or less than that of the corresponding biased estimator according as

$$\rho \left( \frac{\bar{y}^2}{\bar{x}}, \bar{x} \right) \begin{matrix} \leq \\ > \end{matrix} 0. \quad \dots (2.4)$$

The modification of the procedure of *sampling with equal probability with replacement* which provides an unbiased ratio estimator would be to select one unit with ppx, replace it and then select the rest of the  $(n-1)$  units from the whole population with equal probability with replacement at each draw. With this selection procedure the ratio estimator given in (2.2) is unbiased for estimating the population ratio  $R$ , since the probability of getting a particular sample in this case is

$$P(s) = \frac{1}{N^n} \cdot \frac{n!}{\prod_{i=1}^v \lambda_i!} \cdot \frac{\bar{x}}{\bar{X}}, \quad \dots (2.5)$$

where  $\lambda_i$  is the number of repetitions of the  $i$ -th unit and  $v$  is the number of distinct units in the sample. The sampling variance and an unbiased variance estimator of



ratio estimator in this case would be different from those in the case of sampling with equal probability without replacement. The variance estimator is given by

$$\hat{V}(\hat{R}) = \hat{R}^2 - \frac{\sum_{i=1}^v \lambda_i(\lambda_i - 1)y_i^2 + 2 \sum_{i>j}^v \lambda_i \lambda_j y_i y_j}{n(n-1)\bar{X}\bar{x}}. \quad \dots (2.6)$$

In the case of *sampling with equal probability systematically*, an unbiased ratio estimator could be obtained by considering each unit as made up of  $n$  sub-units with each sub-unit of the  $i$ -th unit having the size  $\frac{X_i}{n}$  and selecting one sub-unit with ppx and the other sub-units in the sample systematically proceeding cyclically with the sub-unit selected first as the random start and  $N$  as the sampling interval. The probability of getting a particular sample  $s$  is given by

$$P(s) = \frac{\bar{x}}{\bar{X}}. \quad \dots (2.7)$$

With this probability scheme the estimator given in (2.2) is an unbiased estimator of  $R$ . The variance of the ratio estimator in this case is different from those of estimators based on equal probability selection with or without replacement. Since the selection is being done systematically in this sampling scheme, it would not be possible to get an unbiased variance estimator of the unbiased ratio estimator from a single sample.

It is to be noted that even if the values of the variate  $x$  coming in the denominator are not known for all the units in the population at the time of selection, it is possible to select one unit with ppx by adopting Lahiri's method provided an upper bound of the values of  $x$  which is not much greater than the maximum value is known. The population total of the variate  $x$  would be necessary only if the population total  $Y$  is to be estimated using  $x$  as the supplementary variate.

### 3. VARYING PROBABILITY SAMPLING

The ratio of an unbiased estimator of  $Y$  to that of  $X$  based on pps sampling scheme, the size being the value of a variate  $x$  related to both the characteristics  $x$  and  $y$ , is known to be biased for the population ratio  $R = Y/X$ . In this section are given the selection procedures which provide unbiased ratio estimators corresponding to the usual biased ratio estimators in the case of sampling with pps with replacement, pps without replacement and pps systematically.

The modification of the *pps with replacement scheme* consists in selecting first one unit with ppx, replacing it and then selecting the rest of the  $(n-1)$  units from the whole population with ppz with replacement. The probability of getting a particular sample  $s$  by this procedure is given by

$$P(s) = \frac{n! \prod_{i=1}^v p_i^{\lambda_i}}{X \prod_{i=1}^v \lambda_i!} \left( \frac{1}{n} \sum_{i=1}^n \frac{x_i}{p_i} \right) \quad \dots (3.1)$$

where  $\lambda_i$  and  $N$  are as in (2.5) and  $p_i = \frac{z_i}{Z}$  where  $Z = \sum_{i=1}^N z_i$ . It may be verified that in this case an unbiased estimator of  $R$  is given by

$$\hat{R} = \frac{\frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}}{\frac{1}{n} \sum_{i=1}^n \frac{x_i}{p_i}}. \quad \dots (3.2)$$

It may be noted that the expression for the unbiased ratio estimator is the same as that of the usual biased ratio estimator in the case of pps with replacement sampling. For in the latter case  $\frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}$  and  $\frac{1}{n} \sum_{i=1}^n \frac{x_i}{p_i}$  are unbiased estimators of  $Y$  and  $X$  respectively. Of course, the variance and variance estimator would be different in the two cases.

If in the above selection procedure the units sampled are not replaced before the next and subsequent draws, we get the modified *pps without replacement scheme* which provides an unbiased ratio estimator. But in practice, it is difficult to compute the estimate as the computations involved are quite heavy except in some special cases. Two such special cases have been considered to illustrate the method.

(i) *ppx and ppz of the remaining ( $n = 2$ ).* The modification of this procedure consists in selecting first one unit with ppx and another unit from the remaining  $(N-1)$  units with ppz ( $z$  being a size other than  $x$ ). The probability of getting a particular sample  $s$  ( $x_1 x_2$ ) is given by

$$P(s) = \frac{x_1}{X} \cdot \frac{p_2}{1-p_1} + \frac{x_2}{X} \cdot \frac{p_1}{1-p_2}. \quad \dots (3.3)$$

It can be seen that this procedure provides the following unbiased estimator of the ratio  $R$ .

$$\hat{R} = \frac{\frac{y_1(1-p_2) + y_2(1-p_1)}{p_1}}{\frac{x_1(1-p_2) + x_2(1-p_1)}{p_2}} \quad \dots (3.4)$$

(ii) *ppx, ppz of the remaining and then equal probabilities.* The modification of this procedure consists in following up the procedure explained for case (i) above by selecting the rest of  $(n-2)$  units with equal probability without replacement from the remaining  $(N-2)$  units in the population. This selection procedure makes the following estimator unbiased for the ratio  $R$

$$\hat{R} = \frac{\sum_{i=1}^n \frac{y_i}{1-p_i} \left( \sum_{j \neq i}^n p_j \right)}{\sum_{i=1}^n \frac{x_i}{1-p_i} \left( \sum_{j \neq i}^n p_j \right)}, \quad \dots (3.5)$$



since in this case the probability of getting a particular sample  $s$  is given by

$$P(s) = \frac{\sum_{i=1}^n \frac{x_i}{1-p_i} \left( \sum_{j \neq i}^n p_j \right)}{X \binom{N-2}{n-2}} \dots (3.6)$$

Before giving the selection procedure which provides an unbiased ratio estimator corresponding to the usual biased ratio estimator in the case of *pps systematic sampling*, the method of sampling with probability proportional to size (say the value of a character  $z$ ) systematically will be briefly explained, since this has not become, as yet, well known. Let  $Z_1 Z_2 \dots Z_N$  be the sizes of the units in the population. Suppose the  $i$ -th unit is made up of  $nZ_i$  sub-units, each having the value  $\frac{Y_i}{nZ_i}$  for the variate  $y$ . The procedure of *pps systematic sampling* consists in selecting a random number from 1 to  $Z \left( = \sum_{i=1}^N Z_i \right)$ . The sub-unit having that number is selected in the sample together with every subsequent  $Z$ -th sub-unit. As the total number of sub-units is  $nZ$  there will be  $n$  sub-units in the sample. Let the sample be  $(z_1, z_2 \dots z_n)$ . An unbiased estimator of the population total  $Y$  is given by

$$\hat{y} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i} \dots (3.7)$$

where  $p_i = \frac{z_i}{Z}$ . It may be noted that this estimator resembles that used in the case of *pps* with replacement sampling. But the variances of these two estimators are different.

Though the expected number of repetitions in a sample for the  $i$ -th unit is  $np_i$  in both the *pps* with replacement scheme and the *pps systematic sampling*, the numbers of possible repetitions in a sample are different. For instance, in *pps* with replacement scheme, the  $i$ -th unit may occur 0, 1, 2 ...  $n$  times in a sample of size  $n$  whereas in *pps systematic sampling* it occurs either  $[np_i]$  or  $[np_i]+1$  times. As the randomisation of the number of repetitions is over a smaller range in the case of *pps systematic sampling* than that in the case of *pps* with replacement, it is expected that the former method is more efficient than the latter. Further the efficiency of the estimator based on this method could be increased appreciably by effecting a suitable arrangement of the units in the population before selection. Being a systematically drawn sample it would not be possible to estimate the sampling variance unbiasedly from a single sample.

The modification of the above method to provide an unbiased estimator of  $R = \frac{Y}{X}$  consists in selecting one sub-unit with probability proportional to  $x_i/p_i$ . Then with that as the random start a systematic sample of  $n$  sub-units is selected proceeding

cyclically with  $Z$  as the sampling interval. The probability of getting a particular sample  $s$  is given by

$$P(s) = \frac{1}{ZX} \cdot \frac{1}{n} \cdot \sum_{i=1}^n \frac{x_i}{p_i} \quad \dots \quad (3.8)$$

This procedure provides the following unbiased estimation of the ratio  $R$

$$\hat{R} = \frac{\frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}}{\frac{1}{n} \sum_{i=1}^n \frac{x_i}{p_i}} \quad \dots \quad (3.9)$$

#### 4. STRATIFIED SAMPLING

Let  $k$  be the number of strata and  $N_i$  and  $n_i$  be the number of units in the population and the sample respectively for the  $i$ -th stratum. For stratified simple random sampling without replacement the modification in the selection procedure for getting an unbiased ratio estimator consists of selecting one unit (say the  $j$ -th unit in the  $i$ -th stratum) from the whole population with ppx,  $(n_i-1)$  units from the remaining  $(N_i-1)$  units in the  $i$ -th stratum and  $n_{i'}$  units from  $N_{i'}$  units of the  $i'$ -th stratum ( $i' \neq i$ ) with equal probability without replacement. The probability of getting a particular sample  $s$  is given by

$$P(s) = \frac{\sum_{i=1}^k N_i \bar{x}_i}{X \prod_{i=1}^k \binom{N_i}{n_i}} \quad \dots \quad (4.1)$$

where  $\bar{x}_i$  is the sample mean in the  $i$ -th stratum for the variate  $x$ . With this procedure an unbiased estimator of the ratio  $R$  is given by

$$\hat{R} = \frac{\sum_{i=1}^k N_i \bar{y}_i}{\sum_{i=1}^k N_i \bar{x}_i} \quad \dots \quad (4.2)$$

An unbiased estimator of the variance of  $\hat{R}$  is given by

$$\begin{aligned} \hat{V}(\hat{R}) = \hat{R}^2 - \frac{1}{X \left( \sum_{i=1}^k N_i \bar{x}_i \right)} & \left[ \sum_{i=1}^k \frac{N_i}{n_i} \sum_{j=1}^n y_{ij}^2 + \sum_{i=1}^k \frac{N_i(N_i-1)}{n_i(n_i-1)} \sum_{\substack{j, j' \\ j \neq j'}}^n y_{ij} y_{ij'} + \right. \\ & \left. + \sum_{i \neq i'}^k \frac{N_i N_{i'}}{n_i n_{i'}} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} y_{ij} y_{i'j'} \right] \quad \dots \quad (4.3) \end{aligned}$$



It may be noted that  $\hat{R}$  resembles the biased combined ratio estimator of  $Y$ . The modifications of sampling schemes with other types of designs in the strata can be given on similar lines with a view to getting unbiased ratio estimators.

### 5. TWO-STAGE SAMPLING

In the case of a two-stage sampling design with equal probability selection without replacement at each stage, the selection procedure for providing an unbiased ratio estimator consists in selecting one second stage unit from the whole population of second stage units with ppx. If this second stage unit is from the  $i$ -th first stage unit, the rest of  $(n_i-1)$  second stage units to be sampled from the  $i$ -th first stage unit are selected from the remaining  $(N_i-1)$  units there with equal probability without replacement. The rest of the sample of  $(n-1)$  first stage units is drawn from the remaining  $(N-1)$  units with equal probability without replacement. From these selected first stage units the required number of second stage units are selected with equal probability without replacement. In this case the probability of getting a particular sample  $s$  is given by

$$P(s) = \frac{\sum_{i=1}^n N_i \bar{x}_i}{\binom{N-1}{n-1} \times \prod_{i=1}^n \binom{N_i}{n_i}} \quad \dots (5.1)$$

where  $\bar{x}_i$  is the sample mean in the  $i$ -th selected first stage unit. This selection procedure provides an unbiased estimator of the ratio  $R = \frac{Y}{X}$ .

$$\hat{R} = \frac{\sum_{i=1}^n N_i \bar{y}_i}{\sum_{i=1}^n N_i \bar{x}_i} \quad \dots (5.2)$$

The above procedure could easily be extended to the case of sampling designs with more than two stages, as is shown in the next section.

### 6. MULTI-STAGE DESIGN

The principle involved in giving a selection procedure which provides an unbiased ratio estimator in the case of multi-stage sampling is the same as in the cases illustrated earlier. That is, one final stage unit is to be selected with ppx and the rest of the units according to some probability scheme. For the sake of simplicity only the probability scheme where the units are selected with equal probability without replacement at each stage is considered here.

One final stage unit is to be selected first from the whole population with ppx and then the rest of the sample units are to be selected from the remaining units

in the universe with equal probability without replacement at each stage. This can be achieved as follows. Suppose there are  $m$  stages. One first stage unit ( $i_1$ -th) is to be selected with ppx and the other  $(n-1)$  units with equal probability without replacement from the remaining  $(N-1)$  units. From the first stage unit selected with ppx, one second stage unit is to be selected with ppx and the rest of  $(n_{i_1}-1)$  units are to be selected from the remaining  $(N_{i_1}-1)$  units with equal probability without replacement. Similarly from the  $j$ -th stage unit selected with ppx one  $(j+1)$ th stage unit is to be selected with ppx and the other  $(n_{i_1 i_2 \dots i_j}-1)$  units are to be selected with equal probability without replacement from the remaining  $(N_{i_1 i_2 \dots i_j}-1)$  units. ( $j = 0, 1, 2, (m-1)$ ). From the first and the subsequent stage units selected with equal probability, the required number of higher stage units are to be selected with equal probability without replacement. The probability of getting a particular sample  $s$  is given by

$$P(s) = \frac{\frac{N}{n} \sum_{i_1=1}^n N_{i_1} \bar{x}_{i_1(m)}}{X \prod_{j=0}^{m-1} \left[ \prod_{i_1=1}^n \prod_{i_2=1}^{n_{i_1}} \dots \prod_{i_j=1}^{n_{i_1 i_2 \dots i_{j-1}}} \left( \frac{N_{i_1 i_2 \dots i_j}}{n_{i_1 i_2 \dots i_j}} \right) \right]} \quad \dots \quad (6.1)$$

$$\text{where, } \bar{x}_{i_1(m)} = \frac{1}{n_{i_1}} \sum_{i_2=1}^{n_{i_1}} \frac{N_{i_1 i_2}}{n_{i_1 i_2}} \dots \sum_{i_{m-1}=1}^{n_{i_1 i_2 \dots i_{m-2}}} \frac{N_{i_1 \dots i_{m-1}}}{n_{i_1 \dots i_{m-1}}} \sum_{i_m=1}^{n_{i_1 i_2 \dots i_{m-1}}} x_{i_1 i_2 \dots i_m},$$

$x_{i_1 i_2 \dots i_m}$  being the value of a typical final stage unit. In this case an unbiased estimator of

$$\hat{R} = \frac{\sum_{i_1=1}^n N_{i_1} \bar{y}_{i_1(m)}}{\sum_{i_1=1}^n N_{i_1} \bar{x}_{i_1(m)}}, \quad \dots \quad (6.2)$$

where,  $\bar{y}_{i_1(m)}$  has an interpretation similar to that of  $\bar{x}_{i_1(m)}$ .

## 7. A GENERALISED ESTIMATION PROCEDURE

In this section a generalised procedure for estimating unbiasedly certain types of parameters applicable to a large number of sampling designs is given. For the sake of generality it has become necessary to use some notations which are explained below with suitable examples.

Let  $\mathcal{X}$  denote a population of finite number of units, say, a universe of  $N$  units  $u_1 u_2 \dots u_N$  and  $A$  the class of sets  $\alpha$  whose elements belong to  $\mathcal{X}$ . In such a set the same unit may or may not occur more than once. The class of all point sets and the class of all pairs of units belonging to  $\mathcal{X}$  are examples of the class  $A$ .



Let the population parameter  $F$  be expressible as

$$F = \sum_{\alpha \in A} f(\alpha) \quad \dots (7.1)$$

where  $f(\alpha)$  is a single-valued set function defined over the class  $A$  and  $\sum$  stands for summation over all sets  $\alpha$  belonging to the class  $A$ . For example, the population total  $Y$  can be expressed as  $F$  in (7.1) with  $\alpha$  as a point set  $(u_i)$  and  $f(\alpha)$  as  $y_i$  the value of the  $i$ -th unit for the character  $y$ , and  $Y^2$  can be expressed as  $F$  in (7.1) with  $\alpha$  as a set with two units  $\{u_i, u_j\}^*$  and with  $f(\alpha)$  defined as

$$\begin{aligned} f(\alpha) &= 2y_i y_j & \alpha &= \{u_i, u_j\}, & i &\neq j = 1, 2, \dots N \\ &= y_i^2 & \alpha &= \{u_i, u_i\}, & i &= 1, 2, \dots N \end{aligned}$$

Let a sample  $\omega$  be drawn from the population  $\mathcal{Q}$  with probability  $P(\omega)$ . This  $\omega$  again is a set whose elements belong to  $\mathcal{Q}$ . It may be noted that the same unit may or may not occur more than once in  $\omega$ . The class of all such sets will be denoted by  $\Omega$  which will be the total sample space.

It will be possible to estimate the population parameter  $F$  from the sample  $\omega$  only if each  $\omega$  contains at least one set  $\alpha$  and each set  $\alpha$  is contained in at least one  $\omega$ .

An estimator of the parameter  $F$  is given by

$$\hat{F} = \frac{\sum_{\alpha \subset \omega} f(\alpha) \phi(\omega, \alpha)}{P(\omega)} \quad \dots (7.2)$$

where,  $\sum_{\alpha \subset \omega}$  stands for the summation over all sets  $\alpha$  contained in the sample  $\omega$  and  $\phi(\omega, \alpha)$  is a function of  $\omega$  and  $\alpha$ . This estimator will be unbiased if

$$\sum_{\omega \supset \alpha, \alpha'} \phi(\omega, \alpha) = 1$$

where,  $\sum_{\omega \supset \alpha, \alpha'}$  stands for the summation over all samples  $\omega$  which contain  $\alpha$ , since

$$\begin{aligned} E(\hat{F}) &= \sum_{\omega \in \Omega} \sum_{\alpha \subset \omega} f(\alpha) \phi(\omega, \alpha) \\ &= \sum_{\alpha, \alpha' \in A} f(\alpha) f(\alpha') \left\{ \sum_{\omega \supset \alpha, \alpha'} \psi(\omega, \alpha, \alpha') \right\} \end{aligned}$$

An unbiased estimator of the variance of  $F$  is given by

$$\hat{V}(\hat{F}) = \hat{F}^2 - \frac{\sum_{\alpha, \alpha' \subset \omega} f(\alpha) f(\alpha') \psi(\omega, \alpha, \alpha')}{P(\omega)} \quad \dots (7.3)$$

---

\* The curled brackets  $\{ \}$  are used to denote unordered sets, that is,  $\{u_i, u_j\}$  and  $\{u_j, u_i\}$  are the same.

where,  $\sum_{\alpha, \alpha' \subset \omega}$  stands for the summation over all pairs  $\{\alpha, \alpha'\}$  contained in the sample  $\omega$  and  $\psi(\omega, \alpha, \alpha')$  is a function of  $\omega$  and the pair  $(\alpha, \alpha')$  such that

$$\sum_{\omega \supset \alpha, \alpha'} \psi(\omega, \alpha, \alpha') = 1$$

where,  $\sum_{\omega \supset \alpha, \alpha'}$  stands for the summation over all the samples containing the pair of sets  $(\alpha, \alpha')$ , since

$$\begin{aligned} E \left[ \frac{\sum_{\alpha, \alpha' \subset \omega} f(\alpha) f(\alpha') \psi(\omega, \alpha, \alpha')}{P(\omega)} \right] \\ = \sum_{\alpha, \alpha' \in A} f(\alpha) f(\alpha') \sum_{\omega \supset \alpha, \alpha'} \psi(\omega, \alpha, \alpha') = F^2 \end{aligned}$$

The case where  $\phi(\omega, \alpha)$  is taken as  $P(\omega/\alpha)$ , the conditional probability of getting the sample  $\omega$  given that the set  $\alpha$  has been selected first is of interest as in that case it is possible to verify that for many of the designs in general use, this estimator

$$\hat{F} = \frac{\sum_{\alpha \subset \omega} f(\alpha) P(\omega/\alpha)}{P(\omega)} \quad \dots \quad (7.4)$$

reduces to the usual estimators of the parameter. An unbiased estimator of its variance is given by

$$\hat{V}(\hat{F}) = \hat{F}^2 - \frac{\sum_{\alpha, \alpha' \subset \omega} f(\alpha) f(\alpha') P(\omega/\alpha \cup \alpha')}{P(\omega)} \quad \dots \quad (7.5)$$

where,  $P(\omega/\alpha \cup \alpha')$  is the conditional probability of getting the sample  $\omega$  given that the units in the union of the two sets  $\alpha$  and  $\alpha'$  have been selected first. The above variance estimator may take negative values.

An estimator of the variance is possible only if every set  $(\alpha \cup \alpha')$  is contained in at least one  $\omega$  and every  $\omega$  contains at least one set  $(\alpha \cup \alpha')$ .

## 8. UNBIASED RATIO ESTIMATOR

The above estimation procedure, an estimator for the ratio  $R$  of two parameters  $F$  and  $G$  which can be expressed as

$$\begin{aligned} F &= \sum_{\alpha \in A} f(\alpha) \\ G(\alpha) &= \sum_{\alpha \in A} g(\alpha) \end{aligned}$$

where  $g(\alpha)$  is another single valued set function defined over the class  $A$  is given by

$$\hat{R} = \frac{\sum_{\alpha \subset \omega} f(\alpha) \phi(\omega, \alpha)}{\sum_{\alpha \subset \omega} g(\alpha) \phi(\omega, \alpha)} \quad \dots \quad (8.1)$$



This estimator will be unbiased if

$$P(\omega) = \frac{\sum_{\alpha \subset \omega} g(\alpha) \phi(\omega, \alpha)}{\sum_{\alpha \in A} g(\alpha)} \quad \dots \quad (8.2)$$

since

$$E(\hat{R}) = \sum_{\omega \in \Omega} \frac{\sum_{\alpha \subset \omega} f(\alpha) \phi(\omega, \alpha)}{\sum_{\alpha \subset \omega} g(\alpha) \phi(\omega, \alpha)} P(\omega).$$

If  $g(\alpha)$ 's are either all positive or all negative the above form of  $P(\omega)$  can be obtained by first selecting a set  $\alpha$  with probability proportional to  $g(\alpha)$  and then drawing the rest of the units with some probability scheme. In this case the probability of getting  $\omega$  is

$$P(\omega) = \frac{\sum_{\alpha \subset \omega} g(\alpha) P(\omega/\alpha)}{\sum_{\alpha \in A} g(\alpha)}$$

This shows that if in the general case  $\phi(\omega, \alpha)$  is taken as  $P(\omega/\alpha)$ , the estimator given in (8.1) becomes unbiased for the ratio

$$\hat{R} = \frac{\sum_{\alpha \subset \omega} f(\alpha) P(\omega/\alpha)}{\sum_{\alpha \subset \omega} g(\alpha) P(\omega/\alpha)} \quad \dots \quad (8.3)$$

An unbiased ratio estimator of  $F$  is given by  $\hat{F} = \hat{R} \cdot G$ . ... (8.4)

If  $P(\omega/\alpha)$  is independent of the set  $\alpha$ , the estimator becomes

$$\hat{R} = \frac{\sum_{\alpha \subset \omega} f(\alpha)}{\sum_{\alpha \subset \omega} g(\alpha)} \quad \dots \quad (8.5)$$

and if  $\frac{P(\omega/\alpha)}{h(\alpha)}$  is independent of the set  $\alpha$ ,

$$\hat{R} = \frac{\sum_{\alpha \subset \omega} f(\alpha) h(\alpha)}{\sum_{\alpha \subset \omega} g(\alpha) h(\alpha)} \quad \dots \quad (8.6)$$

An unbiased estimator of the variance of  $\hat{R}$  is given by

$$\hat{V}(\hat{R}) = \hat{R}^2 - \frac{\sum_{\alpha, \alpha' \subset \omega} f(\alpha) f(\alpha') P(\omega/\alpha \cup \alpha')}{G \sum_{\alpha \subset \omega} g(\alpha) P(\omega/\alpha)} \quad \dots \quad (8.7)$$

# SOME SAMPLING SYSTEMS PROVIDING UNBIASED RATIO ESTIMATORS

It is possible that  $F$  and  $G$  can be expressed as sums of set functions defined over more than one class of sets, that is,

$$F = \sum_{\alpha \in A} f_1(\alpha) = \sum_{\alpha' \in A'} f_2(\alpha') = \dots\dots$$

$$G = \sum_{\alpha \in A} f_1(\alpha) = \sum_{\alpha' \in A'} f_2(\alpha') = \dots\dots$$

For each such expression we can give a sampling procedure providing an unbiased ratio estimator. From the point of view of operational convenience, it is preferable to take that class of sets which contains the smaller sets. The size of a set is judged by the number of units it contains. Two examples are given to illustrate the point.

(a) The population total  $X$  can be expressed in the following two ways

$$X = \sum_{i=1}^N X_i = S \frac{\sum_{i=1}^n x_i}{\binom{N-1}{n-1}}$$

where  $S$  stands for the summation over all sets of  $n$  distinct units. In this case the former is to be preferred to the latter because in the former only one unit is to be selected with ppx whereas in the latter case  $n$  units are to be drawn with probability proportional to their total size.

(b) The population variance  $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{X})^2$  where  $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$  can

again be expressed in the following two ways

$$\begin{aligned} \sigma^2 &= \frac{1}{N} \cdot \frac{N-1}{n-1} \cdot \frac{1}{\binom{N}{n}} \cdot S \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{N^2} \cdot S'(x_i - x_j)^2. \end{aligned}$$

Where  $S$  stands for the summation over all sets of  $n$  units,  $\bar{x}$  is the sample mean and  $S'$  stands for the summation over all sets of two units. Here the latter is to be preferred to the former because in the latter case only two units as compared to  $n$  units in the former case are to be selected with probability proportional to their measure of size.



It may be verified that all the cases discussed in earlier sections are particular cases of the generalised unbiased ratio estimator considered here. The procedure explained above will be illustrated by applying it to the question of getting an unbiased estimator of the regression coefficient.

### 9. REGRESSION COEFFICIENT

$$\beta = \frac{\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^N (X_i - \bar{X})^2} \quad \dots (9.1)$$

The numerator and the denominator can be expressed as follows

$$F = \frac{1}{N} \sum_{i=1}^N \sum_{j>i}^N (Y_i - Y_j)(X_i - X_j)$$

$$G = \frac{1}{N} \sum_{i=1}^N \sum_{j>i}^N (X_i - X_j)^2$$

Thus the parameters  $F$  and  $G$  are sums of set functions defined over the class of sets containing only two elements. In the terminology of section 8,

$$f(\alpha) = \frac{1}{N} (Y_i - Y_j)(X_i - X_j)$$

$$g(\alpha) = \frac{1}{N} (X_i - X_j)^2$$

where  $\alpha$  is a set containing two elements.

The selection procedure consists in selecting a pair of units with probability proportional to  $(X_i - X_j)^2$  and the rest  $(n-2)$  units with equal probability without replacement from the remaining  $(N-2)$  units. The conditional probability of getting the sample  $\omega$  given that the pair  $(i, j)$  is selected first is

$$P(\omega/ij) = \frac{1}{\binom{N-2}{n-2}} \quad \dots (9.2)$$

This is independent of the pair of units selected first.

Hence an unbiased estimator of  $\beta$  is given by

$$\hat{\beta} = \frac{\sum_{i=1}^n \sum_{j>i}^n (y_i - y_j)(x_i - x_j)}{\sum_{i=1}^n \sum_{j>i}^n (x_i - x_j)^2}$$

i.e.

$$\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means. The variance and the variance estimator can be got by referring to section 8.

For selecting one pair of units, with probability proportional to  $(X_i - X_j)^2$ , the following procedure may be adopted:

- (i) two random numbers should be selected from 1 to  $N$ . (say,  $ij$ );
- (ii) a pair of random numbers should be selected from 1 to  $\text{Max. } |X_i - X_j|$  ( $= \text{range of } x$ );
- (iii) if both these numbers are less than or equal to  $|X_i - X_j|$  the pair  $(i, j)$  is accepted, otherwise it is rejected;
- (iv) if a pair is rejected, the operation is to be repeated starting from (1).

It may be noted that unbiased estimators of  $\rho^2$  (square of the correlation coefficient of  $x$  and  $y$ ) and  $\beta_2 \left( = \frac{\mu_4}{\mu_2^2} \right)$  can be got by selecting a set of four elements first with suitable probabilities and the rest with any probability scheme.

## 10. TWO-PHASE SAMPLING

For estimating the parameter  $F$  unbiasedly using a ratio estimator with  $g(\alpha)$  as supplementary information, it is necessary to know the value of  $G$ . If the value of  $G$  is not known in advance and if it is easier and less costly to observe  $g(\alpha)$  than  $f(\alpha)$ , then a two-phase design may be used to get an unbiased ratio estimator of  $F$ .

The procedure consists in selecting a large sample  $S$  from the whole population with some probability scheme and observing the value of  $g(\alpha)$  for all sets  $\alpha \in A$  which are contained in  $S$ . A sample  $\omega$  is drawn from  $S$  by first selecting a set  $\alpha \in A$  with probability proportional to  $g(\alpha)P(S/\alpha)$  and then selecting the rest of the sample with some probability scheme. In this case an unbiased ratio estimator of  $F$  is given by

$$\hat{F} = \frac{\sum_{\alpha \subset \omega} f(\alpha)P(S/\alpha).P(\omega/S, \alpha)}{\sum_{\alpha \subset \omega} g(\alpha)P(S/\alpha)P(\omega/S, \alpha)} \times \frac{\sum_{\alpha \subset \omega} g(\alpha)P(S/\alpha)}{P(S)}$$



where  $P(\omega/S, \alpha)$  denotes the probability of selecting the sample  $\omega$  from  $S$  given that  $\alpha$  was selected first in the process. This probability refers to the second-phase sampling.

The authors wish to thank Prof. D. B. Lahiri and Dr. D. Basu for their constant encouragement and advice.

#### REFERENCES

- DES RAJ (1954): Ratio estimation in sampling with equal and unequal probabilities. *Jour. Ind. Soc. Agric. Stat.*, 6, No. 2, 127-138.
- LAHIRI, D. B. (1951): A method of sample selection providing unbiased ratio estimates. *Bull. Int. Stat. Inst.*, 33, part 2, 133-140.
- MIDZUNO, H. (1950): An outline of the theory of sampling systems. *Ann. Inst. Stat. Math.*, 1, 149-156.
- (1952): On the sampling system with probability proportional to sum sizes. (*Ibid*) 3, 99-107.
- SEN, A. R. (1952): Present status of probability sampling and its use in the estimation of characteristics, (abstract). *Econometrica*, 20, 103.

*Paper received : October, 1956.*

*Revised : March, 1959.*

# TABLES FOR SOME SMALL SAMPLE TESTS OF SIGNIFICANCE FOR POISSON DISTRIBUTIONS AND $2 \times 3$ CONTINGENCY TABLES

By I. M. CHAKRAVARTI

and

C. RADHAKRISHNA RAO

*Indian Statistical Institute, Calcutta*

*SUMMARY.* Extended tables of critical values for a level of significance  $\alpha \leq 0.05$  are provided for the variance and likelihood tests for homogeneity, goodness of fit  $\chi^2$  and likelihood tests, and deviation in the zero frequency for samples from a Poisson distribution. A table of critical values for the variance test of homogeneity of samples from truncated Poisson distribution is also given.

For the binomial population, the variance and likelihood tests for homogeneity have been given and tables of critical values worked out, when there are three independent samples of the same size.

Illustrations have been given explaining the use of the various tables.

## 0. INTRODUCTION

In an earlier paper, (Rao and Chakravarti, 1956) exact tests were provided for (i) homogeneity of samples and goodness of fit for Poisson and truncated Poisson populations, and (ii) deviation in the 'zero frequency' for Poisson and binomial populations.

Tables of critical values were given for a level of significance  $\alpha \leq .05$ , for the variance and likelihood tests for homogeneity, goodness of fit  $\chi^2$  and likelihood tests, and deviation in the zero frequency for samples from a Poisson distribution.

These tables have been now extended. A table of critical values of the variance test for homogeneity of samples from truncated Poisson distribution is also provided. Variance and likelihood tests for homogeneity are derived for the binomial population and tables are given for both these tests when there are three independent samples of the same size.



# 1. TEST CRITERIA

1.1. *Poisson population.* Let  $x_1, x_2, \dots, x_f$  be  $f$  observations from Poisson populations; and  $f_r$  denote the frequency of the variate value  $r$  in the sample. Denoting  $\Sigma f_r = f, \Sigma x_i = T$ , the statistics proposed as test criteria were

$\Sigma x_i^2$  : variance test for homogeneity;

$\Sigma x_i \log_e x_i$  : likelihood test for homogeneity;

$\Sigma f_r \log_e (r! f_r)$  : likelihood test for goodness of fit;

$f_0$  : frequency of the zero class for testing excess in that frequency.

The conditional distributions of these statistics given  $T$  and  $f$  were derived.

Tables of critical values of these statistics have been now extended to cover the cases of  $T = 11, 12$ , for  $f = 3(1) 10(10) 100$ .

1.2. *Truncated Poisson.* The analogue of the variance test for homogeneity in the truncated case was found to be

$$\Sigma(x_i - \bar{x})^2 \div \bar{x}(1 + m - \bar{x}) \quad \dots (1.2.1)$$

where  $T = \sum_{i=1}^{f'} x_i = f' \bar{x}$ ,  $\bar{x} = m/(1 - e^{-m})$  and  $x_1, x_2, \dots, x_{f'}$  are  $f'$  observations from a truncated population. The conditional distribution of the statistic (1.2.1) for fixed  $\Sigma x_i$  and  $f'$  is the same as that of  $\Sigma x_i^2$ ; so one can use  $\Sigma x_i^2$  instead. Critical values of this statistic are given for  $f' = 3(1) 9$ ,  $T = 8(1) 12$ .

1.3. *Binomial population.* If  $f$  sets of  $s$  trials are made with probability  $p$  of success and  $x_1, x_2, \dots, x_f$  denote the number of successes in the different sets, then the conditional probability of  $x_1, x_2, \dots, x_f$  given  $\Sigma x_i = T, f$  and  $s$  is

$$\frac{T!(S-T)!}{S!} \prod_{i=1}^f \binom{s}{x_i}$$

where  $S = sf$ .

The statistics which are analogues of variance test and likelihood test for homogeneity are

$$s \Sigma (x_i - \bar{x})^2 / \bar{x}(s - \bar{x}) \text{ and } \Sigma x_i \log_e x_i + \Sigma (s - x_i) \log_e (s - x_i)$$

respectively, where  $\bar{x} = T/f$ . For the former, one may use  $\Sigma x_i^2$ , since the conditional distributions are being considered.

Tables of critical values are provided for these statistics for  $f = 3, s = 3(1)10$ ,  $T = 3(1)26$ .

## SOME SMALL SAMPLE TESTS

To facilitate the computation of likelihood statistics an auxiliary table of  $n \log_e n$  for  $n = 1(1) 100$  is also given.

### 2. USE OF TABLES

2.1. *Critical values.* Below each critical value is recorded the exact level of significance. When this is too low, the next lower value of the criterion is also recorded with the corresponding probability level if it is not much above 5 percent. If the observed value of the statistic is equal to or greater than the tabulated value, then the null hypothesis is rejected at the level of significance indicated.

2.2. Examples : (a) *Poisson distribution.*

no. of accidents	0	1	2	3	total
frequency	32	5	2	1	40
$T = 0 \times 32 + 1 \times 5 + 2 \times 2 + 3 \times 1 = 12.$					

(i) Homogeneity (variance test, Table 1).

$$\sum x_i^2 = \sum r^2 f_r = 1^2 \times 5 + 2^2 \times 2 + 3^2 \times 1 = 22.$$

This value being equal to the critical value for  $T = 12$ ,  $f = 40$ , the null hypothesis of homogeneity is rejected.

(ii) Homogeneity (likelihood test, Table 2).

$$\sum x_i \log_e x_i = \sum (r \log_e r) f_r = 6.068.$$

Critical value of this criterion for  $t = 12$ ,  $f = 40$  is 5.5.

The computed value being greater than this, the samples cannot be regarded as homogeneous.

(iii) Goodness of fit (likelihood test, Table 3).

$$\sum f_r \log_e (f_r r!) = \sum f_r \log_e f_r + \sum f_r \log_e r! = 123.525.$$

The observed value is close to the critical value of 124.76 thus indicating departure from the null hypothesis though not significantly so. The goodness of fit test is, probably, not very sensitive to particular departures from the null hypothesis.

(iv) Test for the deviation in 'zero-frequency' (Table 4).

Frequency of 'zero' = 32.

Since this value is equal to the critical value, a significant excess in the frequency of the zero class is indicated.

(b) *Truncated Poisson.* Homogeneity (variance test, Table 7).



In a study of a rare abnormality in children, 5 families of size 8 reported the following number of abnormal children, 6, 3, 1, 1, 1. Can these samples be regarded as homogeneous ?

Here  $\sum x_i^2 = 6^2 + 3^2 + 1^2 + 1^2 + 1^2 = 48$ . The corresponding critical value for  $f = 5$  and  $T = 12$  is 46 with a level of significance .05. Since the observed value is larger, the hypothesis of homogeneity is rejected.

(c) *Binomial*. Three different preparations  $A$ ,  $B$  and  $C$  of a drug were to be compared for a certain response in mice. Accordingly, 30 animals matched for age and litter were allotted to the three groups at random and the following data were obtained.

preparations	responding	nonresponding	total
$A$	1	9	10
$B$	3	8	10
$C$	8	2	10

Are the preparations equivalent in producing response ?

(i) Homogeneity (variance test, Table 5).

Here  $T = \sum x_i = 1 + 3 + 8 = 12$ ,  $\sum x_i^2 = 1^2 + 3^2 + 8^2 = 74$ .

The observed value of  $\sum x_i^2$  is greater than the corresponding critical value 66 for  $T = 12$ ,  $s = 10$ , at a level of significance 0.03. Hence the test indicates a significant difference in the effects produced by the preparations.

(ii) Homogeneity (likelihood test, Table 6).

$$\begin{aligned}
 & \sum x_i \log_e x_i + \sum (s - x_i) \log_e (s - x_i) \\
 &= 1 \log_e 1 + 3 \log_e 3 + 8 \log_e 8 \\
 &+ 9 \log_e 9 + 8 \log_e 8 + 2 \log_e 2 \\
 &= 57.728.
 \end{aligned}$$

The corresponding critical value being only 52.987 for  $T = 12$ ,  $s = 10$  at a level of significance .03, the test establishes significant differences between the three preparations.

TABLE 1. VARIANCE TEST FOR HOMOGENEITY (POISSON DISTRIBUTION)

Statistic:  $\sum x_i^2 =$  sum of squares of different observations

$f$  = number of observations,  $T$  = total of observations. Values of  $\sum x_i^2$  greater than or equal to the tabulated values are significant at the level indicated in the parentheses.

$f \backslash T$	3	4	5	6	7	8	9	10	20	30	40	50	60	70	80	90	100
3	-	-	9 (.04)	9 (.03)	9 (.02)	9 (.02)	9 (.01)	9 (.01)	9 (.002)	9 (.001)	9 (.001)	9 (.0004)	5 (.05)	5 (.04)	5 (.04)	5 (.03)	5 (.03)
4	16 (.04)	16 (.02)	16 (.01)	16 (.005)	16 (.003)	16 (.002)	16 (.002)	10 <sup>b</sup> (.053)	8 (.02)	8 (.01)	8 (.004)	8 (.003)	8 (.002)	8 (.001)	8 (.001)	8 (.001)	8 (.001)
5	25 (.01)	25 (.004)	17 (.03)	17 (.02)	13 (.04)	13 (.03)	13 (.02)	13 (.01)	11 (.02)	9 (.03)	9 (.01)	9 (.01)	9 (.01)	9 (.005)	9 (.004)	9 (.003)	9 (.002)
6	26 <sup>a</sup> (.053)	26 (.02)	20 (.03)	20 (.01)	18 (.04)	18 (.03)	18 (.02)	18 (.01)	12 (.05)	12 (.02)	10 (.04)	10 (.02)	10 (.02)	10 (.01)	10 (.01)	10 (.01)	10 (.006)
7	37 (.02)	27 (.05)	25 (.03)	25 (.02)	21 (.04)	21 (.02)	19 (.04)	17 (.04)	15 (.02)	13 (.04)	13 (.02)	11 (.05)	11 (.04)	11 (.03)	11 (.02)	11 (.02)	11 (.01)
8	40 (.03)	32 (.03)	30 (.04)	26 (.05)	24 (.03)	24 (.02)	22 (.05)	22 (.03)	18 (.01)	16 (.02)	14 (.04)	14 (.02)	14 (.02)	12 (.05)	12 (.04)	12 (.03)	12 (.02)
9	51 (.02)	39 (.05)	35 (.03)	31 (.05)	29 (.04)	27 (.04)	25 (.04)	23 (.04)	19 (.03)	17 (.04)	17 (.02)	15 (.04)	15 (.03)	15 (.02)	15 (.01)	13 (.05)	13 (.04)
10	58 (.02)	46 (.04)	40 (.04)	36 (.04)	34 (.04)	32 (.03)	30 (.04)	28 (.04)	22 (.03)	20 (.02)	18 (.03)	18 (.02)	16 (.04)	16 (.03)	16 (.02)	16 (.02)	16 (.01)
11	65 (.04)	53 (.04)	47 (.03)	41 (.04)	39 (.03)	37 (.03)	33 (.05)	33 (.03)	25 (.03)	21 (.04)	21 (.01)	19 (.03)	19 (.02)	17 (.04)	17 (.03)	17 (.03)	17 (.02)
12	74 (.05)	60 (.05)	52 (.04)	48 (.03)	42 (.05)	40 (.04)	38 (.04)	36 (.04)	28 (.02)	24 (.03)	22 (.03)	20 (.05)	20 (.03)	20 (.02)	18 (.05)	18 (.04)	18 (.03)

<sup>a</sup> The next higher value is 36 with probability (.004).<sup>b</sup> The next higher value is 16 with probability (.002).



TABLE 2. LIKELIHOOD TEST FOR HOMOGENEITY (POISSON DISTRIBUTION)

Statistic :  $\sum x_i \log_e x_i$ 

$f$  = number of observations,  $T$  = total of all observations. Values of  $\sum x_i \log_e x_i$  greater than or equal to the tabulated values are significant at the level indicated in the parentheses.

$f \backslash T$	3	4	5	6	7	8	9	10	20	30	40	50	60	70	80	90	100
3	-	-	3.2 (.04)	3.2 (.03)	3.2 (.02)	3.2 (.02)	3.2 (.01)	3.2 (.01)	3.2 (.002)	3.2 (.001)	3.2 (.001)	3.2 (.0004)	1.38 (.05)	1.38 (.04)	1.38 (.04)	1.38 (.03)	1.38 (.03)
4	5.5 (.04)	5.5 (.02)	5.5 (.01)	5.5 (.005)	5.5 (.003)	5.5 (.002)	3.2 <sup>b</sup> (.053)	3.2 (.04)	2.7 (.02)	2.7 (.01)	2.7 (.004)	2.7 (.003)	2.7 (.002)	2.7 (.001)	2.7 (.001)	2.7 (.001)	2.7 (.001)
5	8.0 (.01)	8.0 (.004)	5.5 (.03)	5.5 (.02)	4.6 (.04)	4.6 (.03)	4.6 (.02)	4.6 (.01)	3.2 (.02)	2.7 (.03)	2.7 (.01)	2.7 (.01)	2.7 (.01)	2.7 (.005)	2.7 (.004)	2.7 (.003)	2.7 (.002)
6	8.0 <sup>a</sup> (.053)	8.0 (.02)	6.5 (.04)	6.5 (.02)	5.5 (.04)	5.5 (.03)	5.5 (.02)	5.5 (.01)	3.2 (.05)	3.2 (.02)	2.7 (.04)	2.7 (.02)	2.7 (.02)	2.7 (.01)	2.7 (.01)	2.7 (.01)	2.7 (.006)
7	10.7 (.02)	8.8 (.05)	8.0 (.03)	8.0 (.02)	6.9 (.04)	6.0 (.05)	6.0 (.03)	5.5 (.04)	4.1 (.03)	3.2 (.04)	3.2 (.02)	2.7 (.05)	2.7 (.04)	2.7 (.03)	2.7 (.02)	2.7 (.02)	2.7 (.01)
8	12.1 (.03)	10.7 (.03)	9.4 (.04)	8.8 (.04)	7.9 (.04)	7.9 (.03)	6.9 (.05)	6.5 (.05)	5.5 (.01)	4.1 (.03)	3.2 (.04)	3.2 (.02)	3.2 (.02)	2.7 (.05)	2.7 (.04)	2.7 (.03)	2.7 (.02)
9	14.0 (.04)	12.1 (.04)	10.8 (.05)	9.8 (.04)	9.4 (.03)	8.8 (.03)	8.0 (.04)	7.4 (.05)	5.5 (.03)	4.6 (.04)	4.1 (.03)	3.2 (.04)	3.2 (.03)	3.2 (.02)	3.2 (.01)	2.7 (.05)	2.7 (.04)
10	16.2 (.04)	13.6 (.04)	12.4 (.05)	11.0 (.051)	10.7 (.04)	9.7 (.05)	9.3 (.04)	8.8 (.04)	6.5 (.02)	5.5 (.02)	4.6 (.03)	4.1 (.03)	3.2 (.04)	3.2 (.03)	3.2 (.02)	3.2 (.02)	3.2 (.01)
11	18.7 (.04)	16.0 (.04)	14.3 (.04)	13.5 (.03)	12.1 (.05)	11.0 (.04)	10.7 (.03)	9.8 (.05)	6.5 (.05)	5.5 (.05)	4.6 <sup>c</sup> (.052)	4.6 (.03)	4.1 (.04)	3.2 (.04)	3.2 (.03)	3.2 (.03)	3.2 (.02)
12	21.1 (.03)	18.3 (.04)	16.0 (.04)	14.6 (.05)	13.5 (.04)	12.7 (.04)	12.1 (.04)	11.0 (.05)	7.9 (.04)	6.5 (.02)	5.5 (.04)	4.6 (.05)	4.6 (.03)	4.1 (.04)	3.2 (.05)	3.2 (.04)	3.2 (.03)

<sup>a</sup> The next value is 10.7 with prob. (.004).<sup>b</sup> The next value is 5.5 with prob. (.002).<sup>c</sup> The next value is 5.5 with prob. (.02).

TABLE 3. GOODNESS OF FIT FOR A POISSON DISTRIBUTION (LIKELIHOOD TEST)

Statistic :  $\sum f_r \log_e (f_r r!)$  where  $f_r$  = observed frequency of variate value  $r$

$f$  = number of observations,  $T$  = total of all observations. Values of  $\sum f_r \log_e (f_r r!)$  greater than or equal to the tabulated value are significant at a level\* indicated in the parentheses.

$f \backslash T$	3	4	5	6	7	8	9	10	20	30	40	50	60	70	80	90	100
3	-	5.08 (.06)	7.33 (.04)	9.83 (.03)	12.54 (.02)	15.41 (.02)	18.42 (.01)	21.56 (.01)	57.73 (.002)	99.44 (.001)	144.67 (.001)	186.51 <sup>(e)</sup> (.06)	236.19 (.05)	287.62 (.04)	340.51 (.04)	394.69 (.03)	450.02 (.03)
4	4.56 (.04)	6.47 (.02)	8.72 (.01)	11.22 (.005)	13.92 (.003)	13.52 (.04)	16.39 (.04)	19.40 (.03)	53.81 (.02)	95.09 (.01)	140.02 (.004)	187.60 (.003)	237.29 (.002)	288.71 (.001)	341.61 (.001)	395.79 (.001)	451.11 (.001)
5	6.17 (.01)	4.56 <sup>(a)</sup> (.06)	8.04 (.04)	8.72 (.02)	10.53 (.04)	13.23 (.03)	16.10 (.02)	19.12 (.01)	51.34 (.02)	91.76 (.03)	136.37 (.01)	183.72 (.01)	233.22 (.01)	284.48 (.005)	337.24 (.004)	391.30 (.003)	446.51 (.002)
6	7.96 (.004)	6.17 (.05)	7.16 (.04)	9.41 (.04)	10.92 (.04)	12.61 (.05)	15.31 (.03)	18.18 (.02)	49.44 (.05)	89.79 (.02)	133.16 (.04)	180.27 (.02)	229.57 (.02)	280.67 (.01)	333.29 (.01)	387.23 (.01)	442.33 (.01)
7	6.57 (.02)	6.86 (.02)	7.56 (.03)	9.47 (.02)	10.51 (.03)	12.61 (.52)	15.31 (.02)	17.22 (.04)	48.23 (.03)	87.80 (.04)	131.77 (.02)	177.36 (.05)	226.47 (.03)	277.40 (.03)	329.88 (.02)	383.69 (.02)	438.68 (.01)
8	7.96 (.03)	8.31 <sup>(b)</sup> (.052)	8.72 (.03)	9.70 <sup>(d)</sup> (.051)	11.44 (.05)	13.01 (.04)	15.72 (.03)	17.01 (.05)	47.38 (.04)	86.25 (.03)	129.73 (.04)	176.34 (.02)	225.24 (.02)	274.48 (.05)	326.81 (.04)	380.50 (.03)	435.38 (.02)
9	9.21 (.02)	8.67 (.05)	8.94 (.05)	9.87 (.04)	11.20 (.052)	12.71 (.04)	14.52 (.04)	16.77 (.052)	46.57 (.03)	84.94 (.04)	127.92 (.05)	174.27 (.04)	222.96 (.03)	273.56 (.02)	324.02 <sup>(f)</sup> (.060)	377.58 (.05)	432.35 (.04)
10	10.31 (.04)	9.21 (.03)	9.53 (.051)	10.56 (.04)	11.95 (.04)	13.92 (.04)	15.41 (.05)	17.22 (.05)	46.36 (.04)	84.13 (.02)	126.31 (.04)	172.39 (.05)	220.87 (.04)	271.29 (.03)	323.33 (.02)	376.76 (.02)	431.41 (.01)
11	11.36 (.04)	10.31 (.04)	10.60 (.04)	11.14 (.05)	11.95 (.05)	13.69 (.03)	15.65 (.05)	17.92 (.04)	46.19 (.03)	83.73 (.03)	125.49 (.02)	170.68 (.04)	218.88 (.05)	269.11 (.04)	321.07 (.03)	374.37 (.03)	428.91 (.02)
12	13.31 (.05)	12.39 (.04)	11.66 <sup>(e)</sup> (.052)	12.12 (.04)	13.01 (.04)	14.38 (.05)	16.11 (.04)	17.58 (.04)	45.90 (.04)	82.57 (.03)	124.76 (.02)	169.59 (.02)	217.16 (.04)	267.22 (.03)	318.75 (.05)	372.05 (.04)	426.54 (.03)

\* For levels exceeding 0.05, the next higher value of the statistic, with probability in parentheses, is recorded below.

(a) 8.08 (.004) (b) 8.65 (.01) (c) 11.70 (.04) (d) 9.83 (.03) (e) 192.49 (.000) (f) 325.74 (.01).



TABLE 4. TEST FOR FREQUENCY OF THE 'ZERO' CLASS (POISSON DISTRIBUTION)

Statistic : frequency of the 'zero' class

$f$  = number of observations;  $T$  = total of all observations. Values of frequency of 'zero' class greater than or equal to the tabulated values are significant at a level indicated in the parentheses.

$f \backslash T$	3	4	5	6	7	8	9	10	20	30	40	50	60	70	80	90	100
3	-	-	4 (.04)	5 (.03)	6 (.02)	7 (.02)	8 (.01)	9 (.01)	19 (.002)	29 (.001)	39 (.001)	49 (.0004)	58 (.03)	68 (.04)	78 (.04)	88 (.03)	98 (.03)
4	2 (.04)	3 (.02)	4 (.01)	5 (.005)	6 (.003)	7 (.002)	8 (.002)	9 (.001)	18 (.02)	28 (.008)	38 (.004)	48 (.003)	58 (.002)	68 (.001)	78 (.001)	88 (.001)	98 (.001)
5	2 (.01)	3 (.004)	4 (.002)	5 (.001)	6 (.004)	7 (.03)	8 (.02)	9 (.01)	18 (.01)	27 (.03)	37 (.01)	47 (.01)	57 (.01)	67 (.005)	77 (.004)	87 (.003)	97 (.002)
6	2 (.004)	3 (.001)	3 (.04)	4 (.02)	5 (.01)	6 (.01)	7 (.004)	8 (.003)	17 (.01)	27 (.003)	36 (.04)	46 (.02)	56 (.02)	66 (.01)	76 (.01)	86 (.01)	96 (.006)
7	2 (.001)	2 (.05)	3 (.02)	4 (.01)	5 (.003)	5 (.05)	6 (.03)	7 (.02)	16 (.03)	26 (.01)	36 (.005)	45 (.05)	55 (.03)	65 (.03)	75 (.02)	85 (.02)	95 (.01)
8	2 (.0004)	2 (.03)	3 (.01)	4 (.002)	4 (.04)	4 (.03)	6 (.02)	7 (.01)	16 (.01)	25 (.03)	35 (.01)	45 (.01)	55 (.004)	64 (.05)	74 (.05)	84 (.03)	94 (.02)
9	2 (.000)	2 (.01)	3 (.002)	3 (.04)	4 (.02)	5 (.01)	6 (.004)	6 (.04)	15 (.03)	25 (.01)	34 (.03)	44 (.02)	54 (.01)	64 (.01)	74 (.004)	83 (.05)	93 (.04)
10	1 <sup>a</sup> (.052)	2 (.006)	3 (.001)	3 (.02)	4 (.01)	4 <sup>b</sup> (.056)	5 (.03)	6 (.02)	15 (.01)	24 (.02)	34 (.01)	43 (.03)	53 (.02)	63 (.01)	73 (.01)	83 (.01)	93 (.005)
11	1 (.04)	2 (.003)	2 (.04)	3 (.01)	4 (.003)	4 (.03)	5 (.01)	6 (.01)	14 (.03)	23 (.04)	33 (.02)	43 (.01)	52 (.04)	62 (.02)	72 (.02)	82 (.01)	92 (.009)
12	1 (.02)	2 (.001)	2 (.02)	3 (.005)	3 (.04)	4 (.02)	5 (.01)	5 (.04)	14 (.01)	23 (.01)	32 (.03)	42 (.02)	52 (.01)	61 (.04)	71 (.03)	81 (.02)	91 (.02)

<sup>a</sup> The next value is 2 with prob. (.000).<sup>b</sup> The next value is 5 with prob. (.003).

TABLE 5. VARIANCE TEST FOR HOMOGENEITY OF 3 SAMPLES (BINOMIAL DISTRIBUTION)

Statistic:  $\sum x_i^2 = \text{sum of squares of successes}$ 

$s$  = number of trials;  $T = \sum x_i$  = total number of successes. Values of  $\sum x_i^2$  greater than or equal to the tabulated values are significant at a level\* indicated in the parentheses.

$T$	3	4	5	6	7	8	9	10	11	12	13	14
3	9 (.036)			18 (.036)								
4	9 (.055)	16 (.006)	17 (.030)	20 (.039)	25 (.030)	32 (.006)	33 (.055)					
5		16 (.011)	17 <sup>(a)</sup> (.051)	26 (.006)	27 (.021)	32 (.021)	41 (.006)	42 <sup>(e)</sup> (.051)	51 (.011)			
6		16 (.015)	25 (.002)	26 (.012)	27 (.039)	32 (.036)	41 (.025)	44 (.036)	51 (.039)	62 (.012)	73 (.002)	76 (.015)
7		16 (.018)	25 (.003)	26 (.017)	27 <sup>(b)</sup> (.052)	32 (.049)	41 (.042)	46 (.040)	53 (.040)	62 (.042)	67 (.049)	76 <sup>(f)</sup> (.052)
8		16 (.020)	25 (.004)	26 (.021)	29 (.031)	32 <sup>(c)</sup> (.060)	41 <sup>(d)</sup> (.056)	46 <sup>(d)</sup> (.056)	53 <sup>(d)</sup> (.060)	66 (.014)	69 <sup>(j)</sup> (.060)	78 <sup>(k)</sup> (.056)
9		16 (.022)	25 (.005)	26 (.024)	29 (.036)	34 (.047)	45 (.013)	50 (.024)	57 (.032)	66 (.022)	73 (.035)	82 (.035)
10		16 (.023)	25 (.005)	26 (.027)	29 (.040)	34 <sup>(d)</sup> (.053)	45 (.016)	50 (.030)	57 (.041)	66 (.029)	73 (.046)	82 (.050)
$T$	15	16	17	18	19	20	21	22	23	24	25	26
7	89 (.017)	102 (.003)	107 (.018)									
8	89 <sup>(b)</sup> (.056)	96 <sup>(m)</sup> (.060)	109 (.031)	122 (.021)	137 (.004)	144 (.020)						
9	93 (.022)	102 (.032)	113 (.024)	126 (.013)	133 (.047)	146 (.036)	161 (.024)	178 (.005)	187 (.022)			
10	93 (.034)	102 (.050)	113 (.046)	126 (.029)	137 (.041)	150 (.030)	165 (.016)	174 <sup>(n)</sup> (.053)	189 (.040)	206 (.027)	225 (.005)	236 (.023)

\* For levels exceeding 0.05 the next higher value of the statistic, with probability in parentheses, is recorded below.

(a) 22 (.001), (b) 29 (.025), (c) 34 (.040), (d) 38 (.022), (e) 50 (.001), (f) 78 (.025), (g) 45 (.009), (h) 50 (.018), (i) 57 (.022), (j) 73 (.022), (k) 82 (.018), (l) 93 (.009), (m) 98 (.040), (n) 178 (.022).



TABLE 6. LIKELIHOOD TEST FOR HOMOGENEITY OF 3 SAMPLES (BINOMIAL DISTRIBUTION)

Statistic :  $\sum x_i \log_e x_i + \sum (s - x_i) \log_e (s - x_i)$ 

$s$  = number of trials;  $T = \sum x_i$  = total number of successes. Values of  $\sum x_i \log_e x_i + \sum (s - x_i) \log_e (s - x_i)$  greater than or equal to the tabulated values are significant at a level\* indicated in the parentheses.

$T$	3	4	5	6	7	8	9	10	11	12	13	14
3	9.888 (.036)			9.887 (.036)								
4	14.386 (.055)	16.635 (.006)	14.386 (.030)	13.862 (.039)	14.386 (.030)	16.635 (.006)	14.386 (.055)					
5	21.639 (.011)	19.137 <sup>(c)</sup> (.051)	21.639 (.006)	21.639 (.006)	19.137 (.021)	19.137 (.021)	21.639 (.006)	19.137 (.051)	21.639 (.011)			
6	28.432 (.015)	29.548 (.002)	26.844 (.012)	26.844 (.012)	24.141 (.039)	24.613 (.036)	25.729 (.025)	24.613 (.036)	25.729 (.018)	26.844 (.012)	29.548 (.002)	28.432 (.015)
7	36.083 (.018)	36.676 (.013)	33.805 (.017)	33.805 (.017)	32.488 (.025)	31.303 (.049)	30.934 (.042)	30.342 (.040)	30.342 (.040)	30.934 (.042)	31.303 (.049)	32.488 (.025)
8	44.361 (.020)	44.614 (.004)	41.599 (.021)	41.599 (.021)	40.115 (.031)	38.816 <sup>(e)</sup> (.060)	37.895 <sup>(g)</sup> (.056)	37.101 <sup>(h)</sup> (.056)	36.848 (.060)	38.333 (.014)	36.848 <sup>(i)</sup> (.060)	37.101 <sup>(k)</sup> (.056)
9	53.142 (.022)	53.142 (.005)	50.002 (.024)	50.002 (.024)	48.375 (.036)	47.317 (.047)	46.959 (.033)	46.650 (.024)	45.022 (.032)	44.274 <sup>(l)</sup> (.055)	44.728 (.035)	44.728 (.035)
10	62.347 (.023)	62.146 (.005)	58.895 (.027)	58.895 (.027)	57.142 (.040)	55.845 <sup>(f)</sup> (.053)	55.415 (.038)	54.714 (.030)	52.960 (.041)	52.987 (.029)	52.366 (.046)	51.856 (.050)
$T$	15	16	17	18	19	20	21	22	23	24	25	26
7	33.805 (.017)	36.676 (.003)	36.083 (.018)									
8	37.895 <sup>(a)</sup> (.056)	38.816 <sup>(d)</sup> (.060)	40.115 (.031)	41.599 (.021)	44.614 (.004)	44.361 (.020)						
9	44.274 <sup>(b)</sup> (.055)	45.022 (.032)	46.650 (.024)	46.959 (.033)	47.317 (.047)	48.375 (.036)	50.002 (.024)	53.142 (.005)	53.142 (.022)			
10	52.138 (.034)	51.856 (.050)	52.366 (.046)	52.987 (.029)	52.960 (.041)	54.714 (.030)	55.415 (.038)	55.845 <sup>(m)</sup> (.053)	57.142 (.040)	58.895 (.027)	62.146 (.005)	62.347 (.023)

\* For levels exceeding 0.05, the next higher value of the statistic is recorded at the foot of the table.

(a) 39.068 (.027), (b) 45.235 (.022), (c) 24.142 (.001), (d) 39.321 (.040), (e) 39.321 (.040), (f) 56.037 (.042), (g) 39.068 (.027), (h) 39.321 (.018), (i) 37.895 (.022), (j) 37.895 (.022), (k) 39.321 (.018), (l) 45.235 (.022), (m) 56.037 (.042).

# SOME SMALL SAMPLE TESTS

TABLE 7. VARIANCE TEST FOR HOMOGENEITY  
(TRUNCATED POISSON DISTRIBUTION)

Statistic:  $\sum x_i^2$

$f'$  = number of observations,  $T$  = total of observations. Values of  $\sum x_i^2$  greater than or equal to the tabulated values are significant at a level indicated in the parentheses.

$f' \backslash T$	9	8	7	6	5	4	3
12	24 (.02)	32 (.005)	34 (.03)	36 (.045)	46 (.05)	58 (.04)	72 (.05)
11	—	23 (.03)	31 (.007)	33 (.04)	43 (.02)	47 (.05)	68 (.02)
10	—	—	22 (.04)	30 (.01)	40 (.005)	42 (.04)	54 (.04)
9	—	—	—	21 (.05)	29 (.02)	39 (.01)	51 (.01)
8	—	—	—	—	—	28 (.03)	38 (.03)

TABLE 8. VALUES OF  $n \log_e n$

$n$	$n \log_e n$	$n$	$n \log_e n$	$n$	$n \log_e n$	$n$	$n \log_e n$
1		26	84.7105064	51	200.5231107	76	329.1357384
2	1.3862944	27	88.9875963	52	205.4646672	77	334.4730158
3	3.2958372	28	93.3017232	53	210.4254760	78	339.8232864
4	5.5451776	29	97.6515782	54	215.4051414	79	345.1863841
5	8.0471895	30	102.0359250	55	220.4033260	80	350.5621360
6	10.7505576	31	106.4536032	56	225.4196896	81	355.9503771
7	13.6213700	32	110.9035488	57	230.4539298	82	361.3509908
8	16.6355328	33	115.3847475	58	235.5056940	83	366.7637698
9	19.7750214	34	119.8962570	59	240.5747066	84	372.1886112
10	23.0258510	35	124.4371800	60	245.6606820	85	377.6253520
11	26.3768483	36	129.0066804	61	250.7633018	86	383.0738764
12	29.8188780	37	133.6039623	62	255.8823328	87	388.5340134
13	33.3443435	38	138.2282756	63	261.0174798	88	394.0056472
14	36.9468008	39	142.8789024	64	266.1685184	89	399.4886307
15	40.6207545	40	147.5551800	65	271.3351810	90	404.9828640
16	44.3614208	41	152.2564602	66	276.5172102	91	410.4882145
17	48.1646261	42	156.9821232	67	281.7144042	92	416.0045420
18	52.0266906	43	161.7316086	68	286.9265236	93	421.5317442
19	55.9443410	44	166.5043468	69	292.1533485	94	427.0697206
20	59.9146460	45	171.2998125	70	297.3946570	95	432.6183055
21	63.9349704	46	176.1174998	71	302.6502658	96	438.1774176
22	68.0029350	47	180.9569419	72	307.9199592	97	443.7469573
23	72.1163643	48	185.8176432	73	313.2035435	98	449.3268150
24	76.2732888	49	190.6991947	74	318.5008100	99	454.9168701
25	80.4718950	50	195.6011500	75	323.8116150	100	460.5170200



# REFERENCES

- RAO, C. R. AND CHAKRAVARTI, I. M. (1956) : Some small sample tests of significance for a Poisson distribution. *Biometrics*, **12**, 264.
- Papers not cited in the text :*
- COCHRAN, W. G. (1936) : The  $\chi^2$  distribution for Binomial and Poisson series with small expectations. *Ann. Eugen.*, **7**, 207.
- (1954) : Some methods of strengthening  $\chi^2$  tests. *Biometrics*, **10**, 410.
- DAVID, F. N. AND JOHNSON, N. L. (1952) : The truncated Poisson. *Biometrics*, **8**, 275.
- FISHER, R. A. (1921) : On the mathematical foundations of theoretical statistics. *Philos. Trans.*, **A**, 222, 309.
- (1950) : The significance in deviations from expectations in Poisson series. *Biometrics*, **6**, 17.
- HALDANE, J. B. S. (1937) : The exact value of the moments of the distribution of  $\chi^2$  used as a test of goodness of fit, when expectations are small. *Biometrika*, **29**, 133.
- NEYMAN, J. AND PEARSON, E. S. (1928) : On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, **20A**, 175, 263.
- RAO, C. R. (1952) : *Advanced Statistical Methods in Biometric Research*, John Wiley and Sons, New York.
- STEVENS, W. L. (1937) : Significance of grouping. *Ann. Eugen.*, **8**, 57.
- SUKHATME, P. V. (1938) : On the distribution of  $\chi^2$  in samples of the Poisson series. *J. Roy. Stat. Soc.*, **5**, 75.

*Paper received : December, 1958.*

# EXPECTED VALUES OF MEAN SQUARES IN THE ANALYSIS OF INCOMPLETE BLOCK EXPERIMENTS AND SOME COMMENTS BASED ON THEM

By C. RADHAKRISHNA RAO  
*Indian Statistical Institute, Calcutta*

**SUMMARY.** Reference has been made to the logical status of Fisher's null hypothesis that all varieties under test have the same yield on any experimental plot. The types of departures from the null hypothesis, which the ratio of mean squares of varieties to error can detect with a reasonable chance, have been examined in the case of general incomplete block designs. It appears that the test ignores differences in varieties which are, in some sense, attributable to interaction between blocks and treatments. A study of the consequences of non-random allocation of subsets of varieties to blocks leads to a special property of the BIBD and some PBIBD designs. The effect of random indexing of varieties, i.e., of associating the given varieties with the symbols in which a design is represented, is also considered.

## 1. INTRODUCTION

In earlier papers (Rao, 1947, 1956) on general methods of analysis for incomplete block designs, the author has shown how combined intra and inter-block estimates and the expressions for their variances and covariances can be obtained from the theory of least squares under the hypothesis that *treatment* and *plot* effects are additive. The present paper is intended to clarify some of the points not fully elaborated in the earlier papers. Further, accepting Fisher's null hypothesis that an observed yield of a variety on a particular plot is purely a plot effect independent of the variety,<sup>1</sup> the types of departures from the null hypothesis which the analysis of variance test can detect have been examined. The latter is done by comparing the expected values of the mean squares for varieties and error in the analysis of variance under a general hypothesis that on each plot the varieties have possibly different yields, and plot  $\times$  treatment and block  $\times$  treatment interactions exist. The first attempt in this direction was due to Neyman (1935), who obtained the expectations for Randomized block and Latin square designs. Recently Wilk (1955), Wilk and Kempthorne (1957), and others have considered the two cases treated by Neyman under a more general set up.

The null hypothesis is sometimes stated as the equality of varieties with respect to the total yields over all plots of the experimental area *although* plot  $\times$  treatment and block  $\times$  treatment interactions may exist (Neyman, 1935). Some concern is expressed when it is found that under these conditions the expected mean square for varieties is smaller than that for error implying that the analysis of variance ratio test is not unbiased and probably insensitive.

It may be noted that when the total yields of some varieties over a given area are all equal, there must exist portions of the area over which they must differ if interactions exist. The validity of the null hypothesis that the total yields are the

---

<sup>1</sup> In this paper no distinction is made between variety and the more general term treatment. They have been used synonymously.



same over an experimental area then depends largely on what particular area has been chosen or was available for the experiment. Such a null hypothesis has, therefore, not the same logical status as Fisher's null hypothesis. We may now raise the question as to what type of departures from Fisher's null hypothesis are detectable by the variance ratio test of mean square for varieties to that for error. It turns out, in the cases examined before such as Randomized block and Latin square designs as well as for other designs considered in the present paper, the variance ratio test has only a small chance of detecting overall differences equal to or smaller in magnitude than the block  $\times$  treatment interactions. Differences comparatively larger than the interactions have, however, a reasonable chance of detection. It is, perhaps, a desirable property of the test that it should ignore overall differences of the order attributable, in some sense, to the presence of interactions. The object of the experiment may not be to examine whether differences exist over the experimental area used, but to look for evidence whether the results of the experiment would justify an investigation on a large scale, over a wider area. The variance ratio test seems best suited for this purpose. Indeed, if the experimental area itself is chosen at random from a wider area the ratio of variances provides an unbiased test for examining the differences in varieties over the wider area.

Section 2 of this paper is devoted to a brief restatement of some of the results of the earlier paper (Rao, 1947) to clarify some of the statements made earlier and to explain the new notations used in the present communication.

## 2. RANDOMIZATION ANALYSIS OF INCOMPLETE BLOCK DESIGNS UNDER AN ADDITIVE MODEL

Let us consider an incomplete block design involving  $v$  varieties arranged in  $b$  subsets of  $k$  varieties each, such that every variety is used  $r$  times and any pair of varieties  $g$  and  $h$  occurs in  $\lambda_{gh}$  subsets. The actual layout of the experiment in  $b$  blocks of  $k$  plots is determined by the following randomization procedures.

$R_1$  : The subsets of varieties are assigned to the blocks at random.

$R_2$  : Within each block, the varieties of a subset are assigned to the plots at random.

The null hypothesis specifies that all varieties give the same yield on each plot of the experimental area. We may, however, write the yield of the  $g$ -th variety on the  $j$ -th plot of the  $i$ -th block as

$$\tau_g + x_{ij}, \quad g = 1, \dots, v \quad \dots (2.1)$$

where the parameter  $\tau_g$  is specific for the  $g$ -th variety and  $x_{ij}$  is independent of the treatment and may be considered as a plot effect. With the specification (2.1) the null hypothesis under test is

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_v.$$

# EXPECTED VALUES OF MEAN SQUARES

Let us define, with the usual notations for averages,

$$\sigma_p^2 = \frac{\sum \sum (x_{ij} - \bar{x}_{i.})^2}{b(k-1)}, \quad \sigma_b^2 = \frac{\sum (\bar{x}_{i.} - \bar{x}_{..})^2}{b-1}$$

so that  $\sigma_p^2$  is the inherent average variation between plots within blocks and  $\sigma_b^2$  is the variation between blocks. Consider a particular subset of varieties, say  $\tau_1, \dots, \tau_k$  without loss of generality, and represent the corresponding observed yields by  $x^1, \dots, x^k$ . Then it is easy to see that under the randomization procedures  $R_1$  and  $R_2$ ,

$$E(x^i) = \tau_i + \bar{x}_{..}$$

$$V(x^i) = \frac{k-1}{k} \sigma_p^2 + \frac{b-1}{b} \sigma_b^2$$

$$\text{cov}(x^i x^j) = -\frac{1}{k} \sigma_p^2 + \frac{b-1}{b} \sigma_b^2$$

and there exists an orthogonal transformation

$$B/\sqrt{k} = (x^1 + \dots + x^k)/\sqrt{k} \quad \dots \quad (2.2)$$

$$y_i = b_{i1} x^1 + \dots + b_{ik} x^k, \quad \dots \quad (2.3)$$

$$i = 1, \dots, k-1$$

such that the new variables are all uncorrelated and

$$V(B/\sqrt{k}) = k(b-1)\sigma_b^2/b$$

$$V(y_i) = \sigma_p^2, \quad i = 1, \dots, k-1$$

$$E(y_i) = b_{i1} \tau_1 + \dots + b_{ik} \tau_k.$$

An experiment with  $b$  blocks provides  $b(k-1)$  observations of the type (2.3), which are all uncorrelated, have the same variance  $\sigma_p^2$  and have as their expectations linear functions of the unknown parameters  $\tau$ . Hence the theory of least squares can be used for obtaining the best linear estimates of treatment differences  $\tau_i - \tau_j$ , expressions for variances of estimates and the analysis of variance for testing any set of linear hypotheses. This supplies the theory of intra-block analysis.

If, in addition, we make an orthogonal transformation of the block totals (divided by  $\sqrt{k}$ )

$$z_0 = (B_1 + \dots + B_b)/\sqrt{kb}$$

$$z_i = (c_{i1} B_1 + \dots + c_{ib} B_b)/\sqrt{k} \quad \dots \quad (2.4)$$

$$i = 1, \dots, b-1$$

we find that all  $z$  are uncorrelated,  $V(z_i) = k\sigma_b^2$ ,  $i = 1, \dots, b-1$ , and  $E(z_i)$  is a linear function of the varietal effects. The  $(b-1)$  observations (2.4), each with variance



$k\sigma_b^2$  together with the  $b(k-1)$  observations (2.3), each with variance  $\sigma_p^2$  yield combined intra and inter-block estimates by the use of weighted least square theory. The expressions for variances etc., involving the reciprocal of the variances

$$w = 1/\sigma_p^2 \quad \text{and} \quad w' = 1/k\sigma_b^2 \quad \dots (2.5)$$

are given in the author's earlier papers (Rao, 1947, 1956) where a different interpretation was given to  $\sigma_b^2$ .

One useful result of the earlier papers was the derivation of the formulae for varietal differences, and variances and covariances in such a way that the same expressions can be used both for intra-block analysis and combined intra and inter block analysis by substituting appropriate values for the parameters.

To obtain estimates of  $\sigma_p^2$  and  $\sigma_b^2$ , we use the expectations of mean squares in the analysis of variance of total sum of squares into blocks, varieties and error. Table 1 contains the relevant expectations.

TABLE 1. EXPECTATIONS OF MEAN SQUARES ASSUMING THE SPECIFICATION (2.1)

due to	d.f.	s.s.	expected mean square
blocks (ignoring varieties)	$b-1$	$S_b$	$k\sigma_b^2 + \frac{1}{b-1} \sum_{i < j} \left( \frac{r}{v} - \frac{\lambda_{ij}}{k} \right) (\tau_i - \tau_j)^2$
varieties (eliminating blocks)	$v-1-c$	$S_{v \cdot b}$	$\sigma_p^2 + \frac{1}{k(v-1-c)} \sum_{i < j} \lambda_{ij} (\tau_i - \tau_j)^2$
error	$g$	$S_p$	$\sigma_p^2$
varieties (ignoring blocks)	$v-1$	$S_v$	$\frac{v(k-1)}{v-1} \sigma_p^2 + \frac{v(r-1)}{v-1} \sigma_b^2 + \frac{r}{v(v-1)} \sum_{i < j} (\tau_i - \tau_j)^2$
blocks (eliminating varieties)	$b-1$	$S_{b \cdot v}$	$\frac{v-k(c+1)}{k(b-1)} \sigma_p^2 + \frac{v(r-1)}{k(b-1)} \sigma_b^2$

Note:  $g = bk - b - v + 1 + c$ ,  $c$  = degrees of freedom confounded, which is  $(v-1)$  minus the number of independent varietal contrasts estimable from intra-block analysis. The value of  $c$  is zero when all varietal differences are estimable as in any connected design for varietal trials. The relationship  $S_b + S_{v \cdot b} = S_v + S_{b \cdot v}$  could be utilized in computing the sum of squares  $S_{v \cdot b}$  (or  $S_{b \cdot v}$ ) after obtaining directly  $S_{b \cdot v}$  (or  $S_{v \cdot b}$ ).

The estimates of  $\sigma_p^2$  and  $\sigma_b^2$  are obtained using the mean squares whose expectations are free from varietal differences.

$$\hat{\sigma}_p^2 = S_p \div g$$

$$\hat{\sigma}_b^2 = [kS_{b \cdot v} - (v - kc - k)S_p/g] \div v(r-1).$$

For some designs, known as resolvable designs, it is possible to arrange the subsets into groups forming complete replications. The subsets forming a complete replication are assigned to contiguous blocks so that the variation between replications could be removed from the block differences. Defining  $\sigma_r^2$  as the variance between replications the expectations given in Table 2 are obtained.

# EXPECTED VALUES OF MEAN SQUARES

TABLE 2. EXPECTATIONS OF MEAN SQUARES IN THE FURTHER ANALYSIS OF BLOCK VARIATION, ASSUMING THE SPECIFICATION (2.1)

due to	d.f.	s.s.	expected mean square
replications (ignoring blocks)	$r-1$	$S_r$	$v \sigma_r^2$
blocks (within replications)	$b-r$	$S_{b \cdot r}$	$\frac{v-k(c+1)}{k(b-r)} \sigma_p^2 + \frac{(v-k)(r-1)}{b-r} \sigma_b^2$
blocks (eliminating varieties)	$b-1$	$S_{b \cdot v}$	—

In such a case, the estimate of  $\sigma_b^2$  is

$$\hat{\sigma}_b^2 = [S_{b \cdot r} - (v - kc - k)S_p / gk] \div (v - k)(r - 1).$$

For further details the reader is referred to Rao (1947).

## 3. EXPECTATIONS OF MEAN SQUARES WHEN $R_1$ IS NOT FOLLOWED

We shall consider the situation when the subsets of varieties are not randomly assigned to the blocks but only the varieties within a block are randomized, i.e., only the procedure  $R_2$  of Section 2 is followed. In such a case only intra-block estimation is possible. Considering the set up (2.1) let

$$\sigma_{pi}^2 = \sum_j (x_{ij} - \bar{x}_i.)^2 \div (k-1), \quad i = 1, \dots, b \quad \dots \quad (3.1)$$

be the variation between plots within the  $i$ -th block. The transformed variables  $y_1, \dots, y_{k-1}$ , considered in (2.3) arising out of the subset assigned to the  $i$ -th block have the variance  $\sigma_{pi}^2$ . Since  $\sigma_{pi}^2$  is unknown and cannot be estimated unless some varieties are repeated more than once in each block, it is not possible to adopt the procedure of weighted least squares (of weighting the variables of each block by the reciprocal of the corresponding intra-block variance). Let us, therefore, examine the consequence of ignoring the differences in  $\sigma_{pi}^2$  in estimation and tests of significance of varietal differences.

Let  $v_{ij} \sigma_p^2$  denote the variance of  $(\hat{\tau}_i - \hat{\tau}_j)$ , the intra-block estimate of the difference between the  $i$ -th and  $j$ -th varieties, assuming a common variance  $\sigma_p^2$  for all the blocks. The expressions  $v_{ij}$  for various types of standard designs discussed in literature are known. Let  $\delta_s$  be the sum of  $v_{ij}$  for all possible pairs  $i$  and  $j$  of varieties included in the  $s$ -th block. Thus, if the  $s$ -th block has three varieties designated by 1, 4, 5 then  $\delta_s = v_{14} + v_{15} + v_{45}$ . The expectations of mean squares for varieties and error in the intra-block analysis of variance are given in Table 3.



TABLE 3. EXPECTATIONS OF MEAN SQUARES ASSUMING THE SPECIFICATION (2.1)  
WHEN  $R_1$  IS NOT SATISFIED

due to	d.f.	s.s.	expected mean square
blocks (ignoring varieties)	$b-1$	$S_b$	—
varieties (eliminating blocks)	$v-1$	$S_{v \cdot b}$	$\frac{1}{k(v-1)} (\delta_1 \sigma_{p1}^2 + \dots + \delta_b \sigma_{pb}^2) + \frac{1}{k(v-1)} \sum_{j < i} \lambda_{ij} (\tau_i - \tau_j)^2$
error	$g$	$S_p$	$\frac{b(k-1)}{g} \sigma_p^2 - \frac{1}{kg} (\delta_1 \sigma_{p1}^2 + \dots + \delta_b \sigma_{pb}^2)$

$$\sigma_p^2 = (\sigma_{p1}^2 + \dots + \sigma_{pb}^2) \div b$$

From Table 3, we find that when  $\tau_1 = \dots = \tau_v$ , the expected mean squares for varieties and error are not, in general, the same. By equating the coefficients of  $\sigma_{pi}^2$  in the two expressions, we obtain

$$\frac{\delta_i}{k(v-1)} = \frac{k-1}{g} - \frac{\delta_i}{kg} \text{ or } \delta_i = \frac{k(v-1)}{b}$$

so that the necessary and sufficient condition for the equality of expectations for error and varieties is that  $\delta_i$  are all equal, i.e., the sum of variances of all comparisons within any subset of varieties assigned to a block should be the same, the variances being computed, in the usual way, under the assumption of no difference in intra-block variances.

It is easily seen that this condition is satisfied for the BIBD and PBIBD of the two-associate type with the special values  $\lambda_1 = 1, \lambda_2 = 0$  such as the quasi-factorial. It may be of some interest to characterize an experimental design by the presence or absence of this property. Even if this condition is satisfied, there is the additional difficulty of estimating the exact variance of the estimated difference between two given varieties. The exact variance in such a case is a linear compound of the intra-block variances, which, in general, is not a constant multiple of the expected mean square for error except in the case of complete randomized blocks.

#### 4. EXPECTED MEAN SQUARES UNDER A NON-ADDITIVE MODEL FOR A BIBD

In the general case of a non-additive model the following notations and definitions are used.

- (i)  $x_{ij}^a$  = yield of the  $a$ -th variety in the  $j$ -th plot of the  $i$ -th block
- (ii)  $x_i^a, x_{..}^a, \bar{x}_{i.}^a, \bar{x}_{..}^a$  represent sums and averages over the suffixes replaced by dots
- (iii) Variance between plots within blocks

$$b(k-1)\sigma_p^2(a) = \sum \sum (x_{ij}^a - \bar{x}_{i.}^a)^2, \quad a = 1, \dots, v$$

- (iv) Variance between block-means

$$(b-1)\sigma_b^2(a) = \sum (\bar{x}_{i.}^a - \bar{x}_{..}^a)^2, \quad a = 1, \dots, v$$

# EXPECTED VALUES OF MEAN SQUARES

(v) Interaction variances, plot  $\times$  treatment, within blocks

$$b(k-1) i_p^2(a, c) = \frac{1}{2} \sum \sum \{x_{ij}^a - x_{ij}^c - (\bar{x}_i^a - \bar{x}_i^c)\}^2 \quad a, c = 1, \dots, v$$

(vi) Interaction variances, treatment  $\times$  block, based on mean values of blocks

$$(b-1) i_b^2(a, c) = \frac{1}{2} \sum \{\bar{x}_i^a - \bar{x}_i^c - (\bar{x}_{..}^a - \bar{x}_{..}^c)\}^2 \quad a, c = 1, \dots, v$$

(vii) Average variances summed over the varieties

$$\sigma_p^2 = \sum_a \sigma_p^2(a) \div v, \quad \sigma_b^2 = \sum_a \sigma_b^2(a, a) \div v$$

$$i_p^2 = \sum_a \sum_c i_p^2(a, c) \div v(v-1), \quad i_b^2 = \sum_a \sum_c i_b^2(a, c) \div v(v-1)$$

The expected mean squares in the case of BIBD are given in Table 4.

TABLE 4. EXPECTATIONS OF MEAN SQUARES IN THE ANALYSIS OF A BIBD (NON-ADDITIVE MODEL)

due to	d.f.	s.s.	expected mean square
blocks (eliminating varieties)	$b-1$	$S_{b \cdot v}$	$\frac{v-k}{k(b-1)} \sigma_p^2 + \frac{g}{k(b-1)} i_p^2 - \frac{g}{b-1} i_b^2 + \frac{v(v-1)}{b-1} \sigma_b^2$
varieties (eliminating blocks)	$v-1$	$S_{v \cdot b}$	$\sigma_p^2 - \frac{1}{k} i_p^2 + \frac{v-k}{v-1} i_b^2 + \frac{\lambda v}{k(v-1)} \Sigma(\tau_a - \bar{\tau})^2$
error	$g$	$S_{\bar{p}}$	$\sigma_p^2 - \frac{1}{k} i_p^2 + i_b^2$

The expectations in Table 4 under the general set up enable us to examine the nature of the differences in the varieties which the analysis of variance test can detect. Under the null hypothesis, that all varieties have the same yield on each plot of the experimental area, the expectations of the mean square for error and varieties are the same as shown in Table 1. If this null hypothesis is not true then the expected mean square for varieties exceeds that for error only when

$$\frac{v-k}{v-1} i_b^2 + \frac{\lambda v}{k(v-1)} \Sigma(\tau_a - \bar{\tau})^2 > i_b^2$$

$$\text{or} \quad \sigma_\tau^2 - \frac{1}{b} i_b^2 > 0 \quad \dots \quad (4.1)$$

where  $\sigma_\tau^2 = \Sigma(\tau_a - \bar{\tau})^2 / (v-1)$ , the variance between varietal effects. The relationship (4.1) shows that the analysis of variance test has a reasonable chance of detecting departures from the null hypothesis only when the overall differences in yields exceed a certain magnitude depending on the block  $\times$  treatment interaction.

For examining the hypothesis  $\sigma_\tau^2 = 0$ , although  $i_p^2$  and  $i_b^2$  may not be zero, the analysis of variance test is somewhat conservative as in the case of Randomized block and Latin square designs (Neyman, 1935); the position is, however, slightly



better for a BIBD. The difference between the expected values of mean squares for varieties and error in this case is  $[(k-1)/(v-1)]i_b^2$  which is small when  $k$  is small compared to  $v$ . The corresponding difference for randomized blocks is  $i_b^2$  apart from some difference in the magnitude of  $i_b^2$  itself due to increased block size.

##### 5. EXPECTED MEAN SQUARES IN THE CASE OF A GENERAL INCOMPLETE BLOCK DESIGN

For a BIBD it is seen that under the hypothesis  $\sigma_\tau = 0$ , the expectations of mean squares for varieties and error agree upto terms containing intra-block variances and plot  $\times$  treatment interactions. But this may not be true for a general incomplete block design. Let us consider a block containing variety  $a$  with  $(k-1)$  others denoted by  $1, 2, \dots, k-1$  without loss of generality, and define  $\sigma_p^2 V_a$  as the variance of the least square intra-block estimate of the contrast

$$(k-1)\tau_a - \tau_1 - \dots - \tau_{k-1}$$

under the additive model (2.1) where  $\sigma_p^2$  is the intra-block error. For any standard design for which intra-block estimates are provided, the value of  $V_a$  can be obtained directly by first estimating the contrast and computing its variance. Or, if  $\sigma_p^2 v_{ij}$  stands for the variance of the intra-block estimate of  $(\tau_i - \tau_j)$ , then

$$V_a = \sum k v_{ai} - \sum \sum v_{ij} \quad \dots (5.1)$$

where in the summations  $i$  and  $j$  vary over  $a, 1, \dots, k-1$ .

Let  $\sum_i V_a$  = sum of  $V_a$  for blocks in which the varieties  $i$  and  $a$  occur and

$$\xi_a = \frac{1}{k^2} \sum_a V_a$$

$$2\xi_{ac} = \frac{1}{k^3} (\sum_a V_c + \sum_c V_a - k^2 v_{ac} \lambda_{ac}). \quad \dots (5.2)$$

The expected value of the sum of squares due to varieties eliminating blocks is

$$\sum \xi_a \sigma_p^2(a) + \sum \sum \xi_{ac} i_p^2(a, c) - \sum \sum \left[ \frac{\lambda_{ac}}{bkc} + k\xi_{ac} \right] i_b^2(a, c) + \frac{1}{k} \sum_{i < j} \lambda_{ac} (\tau_a - \tau_c)^2 \quad \dots (5.3)$$

and that due to error is

$$\sum \left[ \frac{r(k-1)}{k} - \xi_a \right] \sigma_p^2(a) - \sum \sum \left[ \xi_{ac} + \frac{\lambda_{ac}}{k^2} \right] i_p^2(a, c) + \sum \sum \left[ \frac{\lambda_{ac}}{k} + k\xi_{ac} \right] i_b^2(a, c) \quad \dots (5.4)$$

The condition for terms involving  $\sigma_p^2(a)$  to be equal in the two expectations (5.3) and (5.4) is

$$\xi_a = \frac{v-1}{v} \text{ (independent of } a). \quad \dots (5.5)$$

## EXPECTED VALUES OF MEAN SQUARES

This is true for BIBD, PBIBD of the quasi-factorial type, and linked block designs (LBD), though not for a general incomplete block design. Similarly for the terms involving plot  $\times$  treatment variances to be equal

$$\xi_{ac} = -\frac{(v-1)\lambda_{ac}}{b(k-1)k^2} \quad \dots \quad (5.6)$$

which is true for BIBD and PBIBD of the quasi-factorial type and not LBD. The condition for terms involving block  $\times$  treatment variances to be equal is

$$\xi_{ac} = -\lambda_{ac} \frac{(v-1)(vr-v+1)}{kb^2(k-1)}. \quad \dots \quad (5.7)$$

It may be examined whether this relation can ever be true and if true, what should be the nature of the design.

The actual expressions for LBD are computed to show the disagreement in the plot  $\times$  treatment interactions terms. The expected sum of squares for varieties is, apart from the term involving varietal differences,

$$\frac{v-1}{v} \Sigma \sigma_p^2(a) - \frac{1}{bk} \Sigma \Sigma \lambda_{ac} \left( \frac{1}{k} + \frac{r-\lambda_{ac}}{r\mu} \right) i_p^2(a, c) + \frac{1}{br\mu} \Sigma \Sigma \lambda_{ac} (r-\lambda_{ac}) i_b^2(a, c) \quad \dots \quad (5.8)$$

and that for error is

$$\frac{g}{v} \Sigma \sigma_p^2(a) - \frac{1}{bk} \Sigma \Sigma \lambda_{ac} \left( \frac{b-1}{k} - \frac{r-\lambda_{ac}}{r\mu} \right) i_p^2(a, c) + \frac{1}{b} \Sigma \Sigma \lambda_{ac} \left( \frac{b-1}{k} - \frac{r-\lambda_{ac}}{r\mu} \right) i_b^2(a, c) \quad \dots \quad (5.9)$$

where  $\mu$  is the number of varieties common to any two blocks in a linked block design.

For the coefficients of  $i_p^2(a, c)$  in (5.8) and (5.9) to be the same,  $\lambda_{ac}$  should be constant, in which case the design is a BIBD. How far the disagreement in the interaction terms can be considered as a drawback of the LB design remains to be examined.

As observed earlier there exist designs for which even the terms involving  $\sigma_p^2(a)$  do not agree and it may be of some interest to obtain a classification of the designs with respect to the terms in which the expectations of mean squares for varieties and error agree.

## 6. RANDOM INDEXING OF VARIETIES

In complete randomized block, Latin square and BIBD designs, the association scheme for the actual varieties is independent of the correspondence set up between the varieties and the symbols in which a design is represented. Thus, in a randomly chosen Latin square of order 4 using the symbols  $A, B, C, D$  it does not matter which of the four varieties is made to correspond with  $A$ , which with  $B$  and so on. But in a design like the quasi-factorial obtained by choosing, as blocks, the rows and columns of a square with  $s^2$  symbols written in the  $s^2$  cells there is the further problem of



assigning the varieties to symbols. In this design differences of varieties chosen to correspond with symbols occurring in the same row or column are estimable with a higher precision than those not in the same row or column. So, if certain comparisons are deemed to be more important than others it may be possible to determine a correspondence which allows the estimation of these comparisons with a higher precision. If no such distinction could be made among the various possible comparisons then we may follow the procedure  $R_3$  stated below.

$R_3$  : Obtain the correspondence between the varieties and the symbols in which a design is represented by randomly permuting the symbols over the varieties.

It is easy to verify that whatever may be the design used, with respect to the reference set generated by the randomization procedures  $R_1$ ,  $R_2$  of Section 2 and  $R_3$  stated above, the following are true.

(i) The variance of the estimated difference between any two varieties is a constant independent of the varieties chosen.

(ii) The expected mean squares for varieties and error are same as those obtained for a BIBD (Table 4).

However, it is not suggested that the procedure  $R_3$  justifies attaching the same precision to all estimated differences from the results of an experiment when the design is not balanced.

*Note* : If each observation in an experiment is subject to an additional independent random error (known as technical error) with variance  $\sigma_e^2$ , the expectations of mean squares in all the Tables of this paper will have the additional term  $\sigma_e^2$  with coefficient unity. If  $\sigma_e^2$  is large, the estimation of variance of block totals presents some difficulty. Some aspects of this will be considered in a subsequent publication.

#### REFERENCES

- NEYMAN, J., IWASKIEWICZ, K. (1935): Statistical problems in agricultural experimentation. Suppl. *J. Roy. Stat. Soc.*, **2**, 107.
- RAO, C. R. (1947): General methods of analysis for incomplete block designs. *J. Amer. Stat. Ass.*, **42**, 541.
- (1956): On the recovery of inter-block information in varietal trials. *Sankhyā*, **17**, 105.
- WILK, M. B. (1955): The randomization analysis of a generalised randomized block design. *Biometrika*, **42**, 70.
- WILK, M. B. AND KEMPTHORNE, O. (1957): Non-additivities in a Latin square design. *J. Amer. Stat. Ass.*, **52**, 218.

*Paper received : March, 1959.*

# SOME REMARKS ON THE MISSING PLOT ANALYSIS

By SUJIT KUMAR MITRA

*Indian Statistical Institute, Calcutta*

**SUMMARY.** The analysis of variance of incomplete data from randomised block and latin square experiments is considered and the expected values, under the null hypothesis, of the treatment and error mean squares are obtained. For simplicity, only the case of a single missing observation is considered. The results could be similarly extended to the case of multiple missing observations in a more general situation where the null hypothesis need not be true.

## 1. INTRODUCTION

It is now recognised that the justification of the customary  $F$ -test in ANOVA for designed experiments has to be sought elsewhere and not in the normality and independence assumptions of the observed random variables (an assumption which is most certainly untrue). Several authors (Neyman (1935); Welch (1937); Pitman (1938); and more recently Kempthorne (1955); Wilk and Kempthorne (1955) among others) investigated the possibility of validating this test as an approximate randomisation test. According to them, the stochastic character of the observed variables is primarily due to the random assignment of the treatments to the experimental units and it is possible (theoretically at least) to write down their joint distribution as soon as the randomisation procedure  $R$  is specified. Consider an experiment involving  $N$  experimental units where it is desired to compare  $t$  treatments. Let  $X_u(k)$  be the (hypothetical) yield of the  $u$ -th experimental unit when it receives treatment  $k$  ( $u = 1, 2, \dots, N$ ,  $k = 1, 2, \dots, t$ ). The treatments are said to be equal in their effects if every plot gives the same yield irrespective of the treatment applied, i.e. if,

$$X_u(1) = X_u(2) = \dots = X_u(t) \text{ for } u = 1, 2, \dots, N. \quad \dots (1.1)$$

Usually however we shall not be interested in establishing such a stringent hypothesis (1.1) and are satisfied in detecting deviations from (1.1) only in so far as they imply, differences in total yields (over all the  $N$  units), i.e., in deviations from

$$\sum_u X_u(1) = \sum_u X_u(2) = \dots = \sum_u X_u(t). \quad \dots (1.2)$$

To what extent this is achieved by the ANOVA  $F$ -test in some classical designs (like the Randomised Block Design, Latin Square etc.) has been rather thoroughly examined by all these authors and for a discussion on this subject the reader is referred to Kempthorne's book (1952). All their researches tend to show that the  $F$ -test is unbiased in a certain sense, namely, if (1.1) be true, both the numerator in  $F$  (the treatment m.s.) and its denominator (the error m.s.) have the same expected value. Conditions under which this is true under (1.2) also are known. The object of the present



paper is to demonstrate (in two simple situations) that this is no longer true with the ANOVA  $F$ -test when the yields on some of the units, in an otherwise well-designed experiment, are missing. For computing the expected values we consider independent repetitions of  $R$ , with the same set of experimental units reporting missing yields each time. They are derived making use of certain known results concerning the average values of mean squares in the analysis of such experiments with complete data.

## 2. RANDOMISED BLOCK DESIGN [ONE PLOT MISSING]

Here the experimental units (plots) are arranged in  $r$  blocks of  $t$  plots each and in each block the  $t$  treatments are assigned to the  $t$  plots completely at random. Let  $X_{ij}(k)$  be the yield of the  $j$ -th plot in block  $i$ , when it receives treatment  $k$  and  $x_{ik}$  the observed yield of treatment  $k$  in block  $i$ . For simplicity of discussion we shall assume that the 1st plot in block 1 is missing which under  $R$  had received treatment  $m$  (itself a random variable), so that  $x_{1m}$  is reported to be missing. In such a case Yates' method of fitting constants (1933) for estimating the missing yield leads to the following estimate for  $x_{1m}$ :

$$\hat{x}_{1m} = \frac{rB'_1 + tT'_m - G'}{(r-1)(t-1)} \quad \dots (2.1)$$

where  $B'_1 =$  total yield for the  $(t-1)$  plots in block 1 for which yields were obtained

$$= \sum_{k \neq m} x_{1k}$$

$T'_m =$  total yield for the  $(r-1)$  plots of treatment  $m$  for which yields were obtained

$$= \sum_{i \neq 1} x_{im}$$

$G' =$  total of all the observed yields,

and the ANOVA table is obtained as follows:

TABLE 2.1. ANOVA FOR A RANDOMISED BLOCK DESIGN  
(ONE PLOT MISSING)

sources of variation	d.f.	sum of squares
treatment	$t-1$	$(T)_m$ (obtained by subtraction)
error	$(r-1)(t-1)-1$	$(E)_m$ (obtained as the error s.s. in the completed data inserting $\hat{x}_{1m}$ for the missing $x_{1m}$ )
treatment + error	$r(t-1)-1$	$(T+E)_m = \sum_{k \neq m} \left( x_{1k} - \frac{B'_1}{t-1} \right)^2 + \sum_{i=2}^r \sum_k \left( x_{ik} - \frac{B_i}{t} \right)^2$
$B_i = \sum_k x_{ik}$		

# SOME REMARKS ON THE MISSING PLOT ANALYSIS

Let  $(T)_c$ ,  $(E)_c$  and  $(T+E)_c$  denote the sums of squares due to Treatment, Error and Treatment + Error respectively computed from the complete data, if  $x_{1m}$  were available. Then the following lemma can be easily established.

Lemma 2.1:

$$(E)_c - (E)_m = \frac{(r-1)(t-1)}{rt} (x_{1m} - \hat{x}_{1m})^2$$

$$(T+E)_c - (T+E)_m = \frac{t-1}{t} \left( x_{1m} - \frac{B'_1}{t-1} \right)^2$$

Hence 
$$\mathcal{E}(E)_m = \mathcal{E}(E)_c - \mathcal{E} \left[ \frac{(r-1)(t-1)}{rt} (x_{1m} - \hat{x}_{1m})^2 \right] \quad \dots (2.2)$$

and 
$$\mathcal{E}(T+E)_m = \mathcal{E}(T+E)_c - \mathcal{E} \left[ \frac{t-1}{t} \left( x_{1m} - \frac{B'_1}{t-1} \right)^2 \right]. \quad \dots (2.3)$$

Let us now assume that (1.1) is true, i.e.

$$X_{ij}(1) = X_{ij}(2) = \dots = X_{ij}(t) = X_{ij} \text{ for all } (ij)$$

and write

$$e_{ij} = X_{ij} - X_{i.} \text{ where } X_{i.} = \frac{1}{t} \sum_j X_{ij}. \quad \dots (2.4)$$

In this case

$$(T+E)_c = \sum_i \sum_j \left( x_{ik} - \frac{B_i}{t} \right)^2 = \sum_i \sum_j e_{ij}^2 = r(t-1)A \text{ (say)}$$

and 
$$\left( x_{1m} - \frac{B'_1}{t-1} \right)^2 = \frac{t^2}{(t-1)^2} \left( x_{1m} - \frac{B_1}{t} \right)^2 = \frac{t^2}{(t-1)^2} e_{11}^2.$$

Hence 
$$(T+E)_m = \sum_i \sum_j e_{ij}^2 - \frac{t}{(t-1)} e_{11}^2 = \mathcal{E}(T+E)_m. \quad \dots (2.5)$$

Also since 
$$\mathcal{E}(T'_m) = \sum_{i=2}^r X_{i.}, \text{ we have}$$

$$\mathcal{E}(\hat{x}_{1m}) = \frac{B'_1}{t-1} \quad \dots (2.6)$$



and hence

$$\begin{aligned}\mathcal{E}(x_{1m} - \hat{x}_{1m})^2 &= \left( x_{1m} - \frac{B'_1}{t-1} \right)^2 + V(x_{1m} - \hat{x}_{1m}) \\ &= \frac{t^2}{(t-1)^2} e_{11}^2 + \frac{t^2}{(r-1)^2(t-1)^2} V(T'_m) \\ &= \frac{t^2}{(t-1)^2} e_{11}^2 + \frac{t}{(r-1)^2(t-1)^2} \sum_{i=2}^r \sum_j e_{ij}^2 \quad \dots \quad (2.7)\end{aligned}$$

It is also known that

$$\mathcal{E}(E)_c = (r-1)(t-1)A. \quad \dots \quad (2.8)$$

Hence

$$\mathcal{E}(E)_m = (r-1)(t-1)A - \frac{(r-1)t}{r(t-1)} e_{11}^2 - \frac{A'}{r}$$

and

$$\mathcal{E}(T)_m = (t-1)A - \frac{t}{r(t-1)} e_{11}^2 + \frac{A'}{r} \quad \dots \quad (2.9)$$

where

$$A' = \frac{1}{(r-1)(t-1)} \sum_{i=2}^r \sum_j e_{ij}^2. \quad \dots \quad (2.10)$$

The expected values of the treatment mean square and the error mean square are shown in Table 2.2.

TABLE 2.2. EXPECTED VALUES OF MEAN SQUARES  
IN TABLE 2.1

sources of variation	d.f.	expected value of mean square
treatment	$t-1$	$A^* + \frac{1}{r(t-1)} (A' - A^*)$
error	$r^2 - r - t$	$A^* + \frac{1}{r(r^2 - r - t)} (A^* - A')$

$$A^* = \frac{1}{r(t-1)-1} \left\{ \sum_i \sum_j e_{ij}^2 - \frac{t}{t-1} e_{11}^2 \right\}$$

Hence the  $F$ -test would be unbiased if and only if  $A^* = A'$ .

Thus if the average error variance in the  $(r-1)$  complete blocks is larger than the error variance of the incomplete block 1, we have  $A' > A^*$ , and then the treatment mean square would have a larger expected value than the error mean square. Consequently we would expect a larger proportion of significant  $F$  values even under the null hypothesis (1.1), than what we normally anticipate at the nominal level of testing.

# SOME REMARKS ON THE MISSING PLOT ANALYSIS

## 3. LATIN SQUARE DESIGN [ONE PLOT MISSING]

Here  $N = t^2$  and the  $t^2$  plots are arranged in  $t$  rows and  $t$  columns. The  $t$  treatments are assigned at random to these plots in such a way that each treatment occurs once in every row and once in every column. The randomisation procedure  $R$  in a Latin square is discussed in Fisher and Yates Tables (1948). Let  $X_{ij}(k)$  be the yield of plot  $(i, j)$  ( $i$ -th row and  $j$ -th column) when it receives treatment  $k$  and  $x_{ik}$  the observed yield of treatment  $k$  in row  $i$ . We shall assume that plot  $(1, 1)$  is missing which under  $R$  had received treatment  $m$  so that  $x_{1m}$  is reported to be missing. Here the estimate for the missing yield (Yates, 1936) is computed as

$$\hat{x}_{1m} = \frac{t R'_1 + t C'_1 + t T'_m - 2G'}{(t-1)(t-2)} \quad \dots (3.1)$$

where

$R'_1$  = total yield for the  $(t-1)$  plots in row 1 for which yields were obtained,  
 $C'_1$  = total yield for the  $(t-1)$  plots in column 1 for which yields were obtained,  
 $T'_m$  = total yield for the  $(t-1)$  plots of treatment  $m$  for which yields were obtained, and  
 $G'$  = total of all the observed yields.

The ANOVA table is obtained as follows :

TABLE 3.1. ANOVA FOR A LATIN SQUARE (ONE PLOT MISSING)

sources of variation	d.f.	sum of squares
treatment	$(t-1)$	$(T)_m$ (obtained by subtraction)
error	$(t-1)(t-2)-1$	$(E)_m$ (obtained as the error s.s. in the completed Latin Square inserting $\hat{x}_{1m}$ for the missing $x_{1m}$ )
treatment+error	$(t-1)^2-1$	$(T+E)_m$ (obtained as the (treatment+error) s.s. in the completed Latin Square inserting $\tilde{x}_{1m}$ for the missing $x_{1m}$ )

$$\tilde{x}_{1m} = \frac{t R'_1 + t C'_1 - G'}{(t-1)^2}$$

Let  $(T)_c$ ,  $(E)_c$  and  $(T+E)_c$  be the sums of squares due to Treatment, Error and Treatment + Error respectively computed from the complete latin square if  $x_{1m}$  were available. Then the following result holds :

Lemma 3.1:

$$(E)_c - (E)_m = \frac{(t-1)(t-2)}{t^2} (x_{1m} - \hat{x}_{1m})^2$$

$$(T+E)_c - (T+E)_m = \frac{(t-1)^2}{t^2} (x_{1m} - \tilde{x}_{1m})^2.$$



Hence 
$$\mathcal{E}(E)_m = \mathcal{E}(E)_c - \mathcal{E} \left[ \frac{(t-1)(t-2)}{t^2} (x_{1m} - \hat{x}_{1m})^2 \right]$$

$$\mathcal{E}(T+E)_m = \mathcal{E}(T+E)_c - \mathcal{E} \left[ \frac{(t-1)^2}{t^2} (x_{1m} - \tilde{x}_{1m})^2 \right]. \quad \dots \quad (3.2)$$

Let us now assume that (1.1) is true, i.e.

$$X_{ij}(1) = X_{ij}(2) = \dots = X_{ij}(t) = X_{ij} \text{ for all } (i, j)$$

and write 
$$e_{ij} = X_{ij} - X_{i.} - X_{.j} + X_{..}, \quad \dots \quad (3.3)$$

where 
$$X_{i.} = \frac{1}{t} \sum_j X_{ij}, \quad X_{.j} = \frac{1}{t} \sum_i X_{ij}, \quad X_{..} = \frac{1}{t} \sum_i X_{i.}.$$

As before it can be easily seen that

$$x_{1m} - \tilde{x}_{1m} = \frac{t^2}{(t-1)^2} e_{11}$$

and that 
$$(T+E)_c = \sum_i \sum_j e_{ij}^2 = (t-1)^2 A \quad (\text{say}).$$

Hence 
$$(T+E)_m = \sum_i \sum_j e_{ij}^2 - \frac{t^2}{(t-1)^2} e_{11}^2 = \mathcal{E}(T+E)_m. \quad \dots \quad (3.4)$$

Also since 
$$\mathcal{E}(T'_m) = \frac{1}{t-1} \sum_{i=2}^t \sum_{j=2}^t X_{ij}, \quad R'_1 = \sum_{j=2}^t X_{1j}, \quad C'_1 = \sum_{i=2}^t X_{i1}$$

and  $G' = \sum_{(ij) \neq (11)} X_{ij}$ , we have

$$\mathcal{E}(\hat{x}_{1m}) = \tilde{x}_{1m} \quad \dots \quad (3.5)$$

and hence 
$$\begin{aligned} \mathcal{E}(x_{1m} - \hat{x}_{1m})^2 &= (x_{1m} - \tilde{x}_{1m})^2 + V(x_{1m} - \hat{x}_{1m}) \\ &= \frac{t^4}{(t-1)^4} e_{11}^2 + \frac{t^2}{(t-1)^2(t-2)^2} V(T'_m). \end{aligned} \quad \dots \quad (3.6)$$

It is known that

$$V(T'_m) = \frac{1}{t-2} \sum_{i=2}^t \sum_{j=2}^t e'_{ij}{}^2 \quad \dots \quad (3.7)$$

where

$$\begin{aligned} e'_{ij} &= X_{ij} - X'_{i.} - X'_{.j} + X'_{..} \\ X'_{i.} &= \frac{1}{t-1} \sum_{j=2}^t X_{ij}, \quad X'_{.j} = \frac{1}{t-1} \sum_{i=2}^t X_{ij}, \quad X'_{..} = \frac{1}{t-1} \sum_{i=2}^t X'_{i.} \end{aligned}$$

# SOME REMARKS ON THE MISSING PLOT ANALYSIS

It is also known that

$$\mathcal{E}(E)_c = (t-1)(t-2)A. \quad \dots \quad (3.8)$$

Hence 
$$\mathcal{E}(E)_m = (t-1)(t-2)A - \frac{t^2(t-2)}{(t-1)^3} e_{11}^2 - \frac{A'}{t-1} \quad \dots \quad (3.9)$$

and 
$$\mathcal{E}(T)_m = (t-1)A - \frac{t^2}{(t-1)^3} e_{11}^2 + \frac{A'}{t-1} \quad \dots \quad (3.10)$$

where 
$$A' = \frac{1}{(t-2)^2} \sum_2^t \sum_2^t e_{ij}'^2.$$

The expected values of the corresponding mean squares are shown in Table 3.2.

TABLE 3.2. EXPECTED VALUES OF MEAN SQUARES  
IN TABLE 3.1

sources of variation	d.f.	expected value of mean square
treatment	$t-1$	$A^* + \frac{1}{(t-1)^2} (A' - A^*)$
error	$t^2 - 3t + 1$	$A^* + \frac{1}{(t^2 - 3t + 1)(t-1)} (A^* - A')$

$$A^* = \frac{1}{(t-1)^2 - 1} \left\{ \sum \sum e_{ij}^2 - \frac{t^2}{(t-1)^2} e_{11}^2 \right\}$$

Hence the  $F$ -test would be unbiased if and only if  $A^* = A'$ .

## 4. CONCLUSION

Unless the exact nature of the process, by which an observation is missed, is known, it is difficult to make any further comments on the missing plot analysis. It may be worthwhile to note in this connection that if one plot is missed at random from all the available plots, the bias disappears in both the cases considered in this paper.

The bias which we noticed in this paper possibly affects the test of equality of treatment effects only in so far as the customary use of the percentage points of the variance ratio distribution for judging the significance of the computed value of  $F$  will either overestimate or underestimate the level of the randomisation test. When the number of missing observations is relatively small, this distortion may be only of minor importance.



## REFERENCES

- FISHER, R. A. AND YATES, F. (1948): *Statistical Tables*, Oliver and Boyd, Edinburg, 3rd edition.
- KEMPTHORNE, O. (1952): *Design and Analysis of Experiments*, John Wiley and Sons, New York.
- (1955): The randomisation theory of experimental inference. *J. Amer. Stat. Ass.*, 50, 946-967.
- NEYMAN, J. (with the co-operation of Iwaskiewicz, K. and Kolodzieczyk, St.) (1935): Statistical problems in agricultural experimentation. *J. Roy. Stat. Soc.*, (Supp.) 2, 107-180.
- PITMAN, E. J. G. (1938): Significance tests which may be applied to samples from any population, III. The analysis of variance test. *Biometrika*, 29, 332-335.
- WELCH, B. L. (1937): On the z-test in randomised blocks and latin squares. *Biometrika*, 29, 21-52.
- WILK, M. B. AND KEMPTHORNE, O. (1955): Derived linear models and their use in the analysis of randomised experiments. *Final Report, Analysis of Variance Project*, Statistical Laboratory, Iowa State College, April 1955.
- YATES, F. (1933): The analysis of replicated experiments when field results are incomplete. *Emp. Jour. Exp. Agri.*, 1, 129-142.
- (1936): Incomplete latin squares. *Jour. Agri. Sc.*, 26, 301-315.

*Paper received : March, 1959.*

# MISCELLANEOUS

## THE USE OF LINEAR ALGEBRA IN DERIVING PRIME POWER FACTORIAL DESIGNS WITH CONFOUNDING AND FRACTIONAL REPLICATION

By NORMAN T. J. BAILEY

*Unit of Biometry, University of Oxford, England*

**SUMMARY.** This paper discusses the derivation of prime power factorial designs, with confounding and fractional replication, using only comparatively elementary results in linear algebra. Some further simplifications and systematization of the standard theory of factorial designs are given. The procedures recommended are also extremely quick in practice, and are very easily understood and acquired by students.

### 1. INTRODUCTION

A detailed account of the design and analysis of  $2^m$ ,  $3^n$  and  $2^m 3^n$  factorial designs was first given by Yates (1937). Then Nair (1938) developed a method of dealing with  $p^n$  designs, where  $p$  is a prime or the power of a prime, based on a theory of interchanges connected with the associated hyper-graeco-latin squares. A more general procedure for constructing  $p^n$  arrangements was subsequently obtained by Bose and Kishen (1940) using the theory of Galois fields and finite geometries, and a more elaborate account of this kind of treatment was later given by Bose (1947). Fisher (1942, 1945) made a considerable advance using methods which, at any rate for  $p$  prime, appealed only to the more elementary properties of groups. These methods were also found suitable by Finney (1945) for the development of fractionally replicated designs. An alternative method of investigation has been given by Rao (1946, 1947, 1950) in terms of combinatorial arrangements called arrays of strength  $d$ . Kempthorne (1947) then made a further simplification and systematization of the technique used by Fisher and Finney, and a more detailed account of this theory has since been given in a standard text-book (Kempthorne, 1952). Additional discussion of these topics appears in Brownlee, Kelly and Loraine (1948) and Brownlee and Loraine (1948), and some useful tables have been presented by Rao (1951).

The present paper reconsiders the Fisher-Finney-Kempthorne approach. But, by relating these methods more explicitly to the standard theory of simultaneous linear equations, it is shown that additional simplification and systematization are possible. We are thus enabled to make a quick *systematic* check that the interactions considered for confounding, or for defining contrasts, do in fact have the properties required; and we can find *automatically* generators for the intra-block subgroup and a single treatment combination from each of the other blocks. The method also gives a simple proof of Fisher's theorem on minimum block size. Although the technique is most readily applied if  $p$  is prime, it is also very convenient if  $p$  is the power of a prime, when we must use the appropriate addition and multiplication tables for the corresponding Galois field. Not only are these procedures very quick in practice, but the writer has found that they are understood and acquired by students more easily than the usual text-book methods, since for the most part only comparatively elementary results in standard linear algebra are used.



## 2. STANDARD DERIVATION OF PRIME POWER FACTORIAL DESIGNS

The present discussion assumes an acquaintance with the following basic ideas, for further details of which Kempthorne (1947, 1952) may be consulted. Suppose we consider a  $p^n$  factorial, where there are  $n$  factors,  $A, B, C, \dots$ , each with  $p$  levels. For the time being we take  $p$  to be a positive prime. Then any interaction component  $A^i B^j C^k \dots$ , with  $p-1$  degrees of freedom, is given by comparisons amongst  $p$  sets of treatment combinations, each set being represented by solutions of one of the  $p$  congruences

$$ix_1 + jx_2 + kx_3 + \dots \equiv 0, 1, \dots, p-1 \pmod{p}. \quad \dots (2.1)$$

The coefficients  $i, j, k, \dots$  must all be restricted to the values  $0, 1, 2, \dots, p-1$ , and in order to obtain a complete and unique enumeration of all the degrees of freedom available we adopt the convention that the first non-zero index to appear in  $A^i B^j C^k \dots$  must be unity. It is easily shown that any two congruences of the type shown in (2.1) differing by at least one coefficient on the left, give rise to two sets of treatment combinations such that any contrast with one degree of freedom from the first set is orthogonal to any contrast with one degree of freedom from the second.

Suppose now we want a design in which  $m$  interaction components are confounded with blocks. Then there are  $p^m$  blocks, the composition of which is given by the solutions of  $p^m$  sets of congruences. Each set of congruences contains  $m$  members. The left hand sides in any set are all different and correspond to the  $m$  interaction components chosen; the right hand sides are a selection, with repetitions allowed, from the numbers  $0, 1, \dots, p-1$ .

Fractional replication is dealt with in a similar fashion by adding to the simultaneous congruences already required for confounding further congruences specifying the defining contrasts. The main difference in these latter congruences is that to each left hand side there corresponds only one number on the right, whose value depends on which fraction of the replicate is being used.

## 3. SOLUTION OF A SYSTEM OF LINEAR EQUATIONS

We shall also require the following results for the solution of a system of simultaneous linear equations. An extended discussion of the basic theory may be found in Chapter 1 of Stoll (1952). Suppose we have a consistent system of  $m$  non-homogeneous linear equations in  $n$  variables  $x_j, j = 1, 2, \dots, n$ , viz.,

$$\sum_{j=1}^n a_{ij} x_j = y_i, \quad i = 1, \dots, m. \quad \dots (3.1)$$

Then it is a straightforward matter to reduce (3.1) to the *echelon* (or canonical) form

$$\sum_{j=1}^{k_i-1} b_{ij} x_j + x_{k_i} = z_i, \quad i = 1, \dots, r \leq m, \quad \dots (3.2)$$

where

$$1 \leq k_1 < k_2 < \dots < k_r \leq n, \quad \dots (3.3)$$

using only the usual elementary operations. The distinguishing characteristics of such a system are (i) the last non-zero coefficient in each equation is unity, (ii) the lengths,  $k_i$ , of the  $r$  linear forms are all different and follow the order of magnitude shown in (3.3), and (iii)  $x_{k_i}$  appears with a non-zero coefficient only in the  $i$ -th equation.

# ON THE DERIVATION OF PRIME POWER FACTORIAL DESIGNS

From (3.2) we obtain immediately the solved form or general solution

$$\left. \begin{aligned} x_{k_i} &= \sum_{j=1}^{n-r} c_{ij} u_j + z_i, & i &= 1, \dots, r; \\ &= u_{i-r}, & i &= r+1, \dots, n, \end{aligned} \right\} \quad \dots \quad (3.4)$$

where the  $u_j$  are arbitrary numbers. We are in fact writing the  $x_{k_i}$ ,  $i = 1, \dots, r$  in terms of the remaining  $n-r$  unknowns to which arbitrary values can be assigned.

We now consider the homogeneous system of equations

$$\sum_{j=1}^n a_{ij} x_j = 0, \quad i = 1, \dots, m, \quad \dots \quad (3.5)$$

given by putting all the  $y_i$  equal to zero in (3.1). The equations corresponding to (3.2) and (3.4) now have all the  $z_i$  zero. Consider the set of solutions given by

$$\left. \begin{aligned} X_1 &= (c_{11}, & c_{21}, & \dots, & c_{r1}, & 1, & 0, & 0, & \dots, & 0), \\ X_2 &= (c_{12}, & c_{22}, & \dots, & c_{r2}, & 0, & 1, & 0, & \dots, & 0), \\ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ X_{n-r} &= (c_{1, n-r}, & c_{2, n-r}, & \dots, & c_{r, n-r}, & 0, & 0, & 0, & \dots, & 1), \end{aligned} \right\} \quad \dots \quad (3.6)$$

where we now take  $x_{k_1}, x_{k_2}, \dots, x_{k_n}$  for the order of the  $x_j$ . It is easily shown that (3.6) is a *basis* of the general solution of (3.5), in the sense that the latter is given by the set,  $X$ , of all

linear combinations like  $\sum_{j=1}^{n-r} u_j X_j$ .

When  $r < n$  it is often useful to employ the theorem that the general solution of (3.1) can be written as  $X + X_0$ , where  $X$  is the general solution of (3.5), as above, and  $X_0$  is one fixed solution of (3.1) and we regard  $X$  and  $X_0$  as linear forms in the  $x_j$ .

The above results hold not only for systems of equations for which the coefficients and unknown  $x_j$  are all real numbers, but also for systems with coefficients, and hence solutions, in any field. This fact permits us to make the application to factorial designs described in the next section.

## 4. APPLICATION TO PRIME POWER FACTORIALS OF TYPE $p^n$

In applying the results of the last section to the type of situation envisaged in section 2, the first remark to be made is that the congruences (2.1) can be replaced by actual equations

$$ix_1 + jx_2 + kx_3 + \dots = 0, 1, \dots, p-1, \quad \dots \quad (4.1)$$

provided that we use an algebra of remainders modulo  $p$ . If  $p$  is prime the system of integers modulo  $p$  constitute a finite field to which all the results of Section 3 may be applied,



Suppose now that we have a  $p^n$  factorial and wish to confound  $m$  interactions with blocks. The required treatment combinations in the  $p^m$  blocks are given by the solutions of the  $p^m$  possible systems of equations. Each system is of the type shown in (3.1) with  $m$  members, and the  $y_i$  are a selection with repetitions from the integers  $0, 1, \dots, p-1$ .

*Checking the interactions to be confounded.* The first step in deriving a design is to check that the interactions to be confounded are in fact linearly independent, and do not automatically entail the confounding of any interactions of order less than some specified number (often three). This is done by reducing the homogeneous system (3.5) to the corresponding echelon system. If  $m$  equations remain then both these and the original system are necessarily independent. All other interactions involved in the confounding must be given by all linear combinations of the rows of coefficients in the echelon system. It can usually be seen at a glance whether the required condition holds. If, for example, we wish no two-factor interaction to be confounded then we first inspect the echelon system to ensure that each equation has more than two non-zero coefficients. Secondly, we need consider only linear combinations of pairs of equations, since combinations of three or more must involve at least three non-zero coefficients. For each pair combined the last non-zero coefficients in each equation provide two non-zero elements, and we have only to ensure that at least one other always survives. This entails examining far fewer elements than the usual procedure.

It should be mentioned that the echelon system technique was employed by Bose (1947) in his more advanced treatment of certain factorial designs. Bose uses the term 'canonical form' instead of echelon system.

*Specifying the intrablock subgroup and other blocks.* Having checked the interactions to be confounded we next determine the composition of the various blocks. The set of treatment combinations appearing in the block containing  $(00\dots 0)$  is given by the solution of the homogeneous system (3.5). This can be done immediately by considering the echelon system as described in the last section. It is often convenient to specify this block, especially if it is fairly large, by means of a basis (with  $n-m$  members), all linear combinations of which give the remaining treatment combinations. Since we are using integers modulo  $p$  the complete solution constitutes a group—the intrablock subgroup—for which the basis is a set of generators.

The composition of any other block, obtained by solving one of the appropriate systems of non-homogeneous equations, is then given by the addition of one fixed solution of the latter to the general solution of the homogeneous system. This is equivalent to the usual rule for deriving blocks other than the intrablock subgroup. It is easily seen that the quickest way to write down one treatment combination for each of these blocks when confounding in a complete replicate is to start with the control  $(00\dots 0)$  appearing in the intrablock subgroup and then let the  $m$  variables corresponding to the  $x_{k_i}$ ,  $i = 1, 2, \dots, m$ , in the echelon system run through all the values  $0, 1, \dots, p-1$ .

With fractional replication on the other hand we must restrict this set of treatment combinations to those satisfying the equations corresponding to the specific defining contrasts adopted. With large fractions this can be done by inspection, but with small fractions it may be more convenient to use a more systematic method as follows. If, when manipulating



# ON THE DERIVATION OF PRIME POWER FACTORIAL DESIGNS

the rows of coefficients in the original simultaneous equations to obtain the echelon system, we perform a similar set of operations on the unit matrix with the same number of rows, a matrix is obtained which when used as a pre-multiplier effects the transformation directly. When this matrix pre-multiplies the matrix given by all admissible sets of quantities appearing on the right of the original equations we obtain a new array whose columns give immediately a possible set of alternative values for the  $x_{k_i}$  when the other variables are all zero.

When using a complete replicate we have of course the original  $p^m$  combinations all over again, and so do not need to go through this procedure at all.

The illustrative example below should make clear how easily the method can be applied in practice.

*Illustrative example.* As an illustration of the foregoing let us consider the example given by Kempthorne [1952, p. 426] of a  $1/9$  replicate of a  $3^7$  factorial confounded in 9 blocks each with 27 treatment combinations. The defining contrasts suggested for the fractional replication were  $ABCD^2E$  and  $CD^2E^2F^2G^2$ , while  $AB^2F^2G$  and  $BCDF$  were in addition to be confounded between blocks. To check these interactions and obtain the intrablock subgroup we consider the systems of equations whose left-hand sides have coefficients given by

$$\begin{bmatrix} 1 & 1 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 0 & 0 & 0 & 2 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix} \quad \dots \quad (4.2)$$

A suitable system is quickly found to be

$$\begin{bmatrix} 2 & 2 & 1 & 0 & 0 & 0 & 0 \\ 2 & 2 & 0 & 2 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \quad \dots \quad (4.3)$$

which is in fact derived from (4.2) by the sequence of linear operations defined by the pre-multiplying matrix

$$\begin{bmatrix} 1 & 1 & 1 & 2 \\ 0 & 2 & 2 & 1 \\ 2 & 2 & 2 & 2 \\ 2 & 2 & 0 & 2 \end{bmatrix} \quad \dots \quad (4.4)$$

(It is easily checked that (4.3) is the product of (4.4) and (4.2) in that order.)

The four interactions originally suggested are clearly independent since we still have four rows in the echelon array (4.3). Examination of the reduced array consisting of the first, second and fourth columns only shows that in no case can a linear combination of two



rows involve less than one non-zero coefficient. The full array thus entails the confounding of no interaction with less than three factors. We accordingly proceed with specifying the defining elements of the design.

A suitable basis for the intrablock subgroup is obtained by adopting in turn the values (100), (010) and (001) for the variables corresponding to the factors  $A$ ,  $B$  and  $D$ , and then solving at sight the homogeneous system of equations whose left-hand sides are represented by (4.3). Thus we first write down the bold-faced numerals in (4.5) below and then fill in the remainder using in turn each of the equations just mentioned. The three generators of the intrablock subgroup are therefore

$$\left. \begin{array}{l} (1 \mathbf{0} \mathbf{1} \mathbf{0} \mathbf{1} \mathbf{2} \mathbf{1}), \\ (0 \mathbf{1} \mathbf{1} \mathbf{0} \mathbf{1} \mathbf{1} \mathbf{2}), \\ (0 \mathbf{0} \mathbf{0} \mathbf{1} \mathbf{1} \mathbf{2} \mathbf{2}). \end{array} \right\} \quad \dots \quad (4.5)$$

If we are to use the fraction containing the 'control' treatment (000 ... 0) then there are nine systems of equations in all, one of which is homogeneous, whose left-hand sides are given by (4.2) and whose right-hand sides may be taken as the columns of the matrix

$$\left[ \begin{array}{cccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 \end{array} \right] \quad \dots \quad (4.6)$$

Pre-multiplication by (4.4) gives

$$\left[ \begin{array}{cccccccc} 0 & 2 & 1 & 1 & 0 & 2 & 2 & 1 \\ 0 & 1 & 2 & 2 & 0 & 1 & 1 & 2 \\ 0 & 2 & 1 & 2 & 1 & 0 & 1 & 0 \\ 0 & 2 & 1 & 0 & 2 & 1 & 0 & 2 \end{array} \right] \quad \dots \quad (4.7)$$

so that nine suitable treatment combinations, one from each block, are

$$\left. \begin{array}{l} (0 \mathbf{0} \mathbf{0} \mathbf{0} \mathbf{0} \mathbf{0} \mathbf{0}), \\ (0 \mathbf{0} \mathbf{2} \mathbf{0} \mathbf{1} \mathbf{2} \mathbf{2}), \\ (0 \mathbf{0} \mathbf{1} \mathbf{0} \mathbf{2} \mathbf{1} \mathbf{1}), \\ (0 \mathbf{0} \mathbf{1} \mathbf{0} \mathbf{2} \mathbf{2} \mathbf{0}), \\ (0 \mathbf{0} \mathbf{0} \mathbf{0} \mathbf{0} \mathbf{1} \mathbf{2}), \\ (0 \mathbf{0} \mathbf{2} \mathbf{0} \mathbf{1} \mathbf{0} \mathbf{1}), \\ (0 \mathbf{0} \mathbf{2} \mathbf{0} \mathbf{1} \mathbf{1} \mathbf{0}), \\ (0 \mathbf{0} \mathbf{1} \mathbf{0} \mathbf{2} \mathbf{0} \mathbf{2}), \\ (0 \mathbf{0} \mathbf{0} \mathbf{0} \mathbf{0} \mathbf{2} \mathbf{1}). \end{array} \right\} \quad \dots \quad (4.8)$$

The whole design can now be written down from the basic defining elements provided by (4.5) and (4.8).

# ON THE DERIVATION OF PRIME POWER FACTORIAL DESIGNS

## 5. FISHER'S THEOREM ON MINIMUM BLOCK SIZE

The theorem proved by Fisher (1942, 1945) on minimum block size is well-known. This states that a  $p^n$  factorial can be arranged so as to confound  $m$  independent interactions between  $p^m$  blocks, each of  $p^{n-m}$  treatment combinations, with no two-factor interaction confounded, provided

$$n \leq \frac{p^{n-m}-1}{p-1} \quad \dots \quad (5.1)$$

When  $p = 2$  this reduces to the requirement that the number of treatment combinations in each block must be greater than the number of factors.

Consider the coefficients in the corresponding echelon system. There are  $m$  columns containing a unit element which is the last in its row. This leaves an array with  $n-m$  columns and  $m$  rows. In order that no two-factor interaction may be confounded it is sufficient if each of these latter rows contains at least two non-zero elements, and if all rows are different subject to none being a multiple of any other. The total number of ways of allotting the numbers  $0, 1, 2, \dots, p-1$  to  $n-m$  places is  $p^{n-m}$ , but we must exclude the single case having all zeros, and the  $(n-m)(p-1)$  cases with just one non-zero element. The remaining arrangements all contain at least two non-zero elements, but to any given arrangement there are  $p-2$  others which are merely multiples. The total number of arrangements available for allocation to the  $m$  rows, such that the required conditions are satisfied, is therefore

$$\frac{p^{n-m}-1-(n-m)(p-1)}{p-1} \quad \dots \quad (5.2)$$

The condition that this expression must be not less than  $m$  yields the required result (5.1) after rearrangement.

## 6. DERIVATION OF A $p^n$ DESIGN FROM A $2^n$ DESIGN

Another point worth commenting on is the way in which a  $p^n$  design may be derived from a  $2^n$ . Fisher (1942) remarked that "the solutions available when  $p = 2$  will be available in general, with the assurance that no interaction will involve fewer factors than the confounding interaction when  $p = 2$ ." To see how this result arises in connexion with the present treatment, we consider the echelon system for the  $m$  homogeneous equations whose solutions give the intrablock subgroup in a  $2^n$  factorial, and the corresponding set of  $n-m$  generators. It is easily seen that if we change to an algebra modulo  $p$  ( $p > 2$ ) then the same set of  $n-m$  generators will satisfy a new echelon system of  $m$  equations which is derived from the first system by multiplying all coefficients, except the last in each row, by  $p-1$ . We now have the essential ingredients of a  $p^n$  design confounding  $m$  interactions, each involving no fewer factors than the original  $2^n$  arrangement, in blocks of  $p^{n-m}$ .



7. EXTENSION TO DESIGNS WITH FACTORS HAVING  $p^s$  LEVELS

So far we have been considering  $p^n$  designs in which each factor has a number of levels,  $p$ , which is prime. When the numbers of levels for the several factors are powers, not all the same, of the same prime, e.g.,  $2^2 \times 4^2$ , the introduction of pseudofactors is usually the most convenient method of treatment as it reduces the design to a standard  $p^n$  form, in this case  $2^6$ . This has however the disadvantage, especially with confounding, that some components of main effects of original factors appear formally as higher order interactions of pseudofactors. If therefore the factors have levels which are all the same power of a prime it may be more convenient to proceed directly.

When dealing with a  $p^n$  factorial we made use of the fact that integers modulo  $p$  constitute a field of  $p$  elements. If the number of levels is the power of a prime,  $p^s$ , we can still use the same basic technique outlined above since it is always possible to construct a field of  $p^s$  elements. The theory of this involves the more advanced properties of Galois fields and has been admirably described in the context of experimental design by Bose (1938). The main complication is that the elements of the field are no longer real numbers, but can be adequately represented by them provided we adopt the appropriate rules for addition and multiplication.

On the whole the remarks of section 2 still apply with the obvious modifications. Thus any interaction component  $A^i B^j C^k \dots$  with  $p^s - 1$  degrees of freedom is given by comparisons amongst  $p^s$  sets of treatment combinations obtained as solutions of equations like

$$ix_1 + jx_2 + kx_3 + \dots = 0, 1, 2, \dots, p^s - 1, \quad \dots (7.1)$$

where now all coefficients and variables are elements of the relevant Galois field, and must obey the appropriate laws of composition. Designs with confounding or fractional replication can be derived from the solution of simultaneous equations as before. The only further material we require is therefore a set of addition and multiplication tables for the small number of Galois fields likely to be required in practice.

*Illustrative example.* A simple illustration should make the simplicity of the above procedure clear. Consider the problem of laying out a  $4^4$  factorial in 16 blocks of 16 units each. The addition and multiplication tables for a Galois field of  $2^2$  elements, 0, 1, 2 and 3, can be taken as

+	0	1	2	3
0	0	1	2	3
1		0	3	2
2			0	1
3				0

·	0	1	2	3
0	0	0	0	0
1		1	2	3
2			3	1
3				2

... (7.2)

# ON THE DERIVATION OF PRIME POWER FACTORIAL DESIGNS

Suppose we decide to try confounding the interactions  $ABC$  and  $BC^2D$ . The coefficients in the homogeneous equations whose solution yields the intrablock subgroup are given by

$$\begin{bmatrix} 1110 \\ 0121 \end{bmatrix} \quad \dots \quad (7.3)$$

Inspection of the first and fourth columns shows that (7.3) can already be taken as an echelon form. We immediately obtain the generators

$$\begin{pmatrix} 1 & \mathbf{1} & \mathbf{0} & 1 \\ 1 & \mathbf{0} & \mathbf{1} & 2 \end{pmatrix}, \quad \dots \quad (7.4)$$

where, as before, the bold-faced numerals are written down first, and the remaining elements are calculated by substitution in the echelon system. Taking all linear combinations of the two generators then gives 13 further treatments, which with (0000) make up the 16 required. One combination from each of the other 15 blocks is obtained by 'ringing the changes' on the numerals in (7.4) which are *not* bold-faced.

We can also introduce fractional replication into factorial designs of the present type by using the device already described in the illustration at the end of Section 4.

## REFERENCES

- BOSE, R. C. (1938): On the application of the properties of Galois fields to the problem of the construction of hyper-graceo-latin squares. *Sankhyā*, **3**, 323-338.
- (1947): Mathematical theory of the symmetrical factorial design. *Sankhyā*, **8**, 107-166.
- BOSE, R. C. AND KISHEN, K. (1940): On the problem of confounding in the general symmetrical factorial design. *Sankhyā*, **5**, 21-36.
- BROWNLEE, K. A., KELLY, B. K. AND LORAINE, P. K. (1948): Fractional replication arrangements for factorial experiments with factors at two levels. *Biometrika*, **35**, 268-276.
- BROWNLEE, K. A. AND LORAINE, P. K. (1948): The relationship between finite groups and completely orthogonal squares, cubes and hyper-cubes. *Biometrika*, **35**, 277-282.
- FINNEY, D. J. (1945): The fractional replication of factorial arrangements. *Ann. Eugen., Lond.*, **12**, 291-301.
- FISHER, R. A. (1942): The theory of confounding in factorial experiments in relation to the theory of groups. *Ann. Eugen., Lond.*, **11**, 341-353.
- (1945): A system of confounding for factors with more than two alternatives, giving completely orthogonal cubes and higher powers. *Ann. Eugen., Lond.*, **12**, 283-290.
- KEMP THORNE, O. (1947): A simple approach to confounding and fractional replication in factorial experiments. *Biometrika*, **34**, 255-272.



- (1952): *The Design and Analysis of Experiments*, John Wiley & sons, New York, Chapman & Hall, London.
- NATH, K. R. (1938): On a method of getting confounded arrangements in the general symmetrical type of experiment. *Sankhyā*, 4, 121-138.
- RAO, C. R. (1946): On hypercubes of strength  $d$  and a system of confounding in factorial experiments. *Bull. Cal. Math. Soc.*, 38, 67-78.
- (1947): Factorial experiments derivable from combinatorial arrangements of arrays. *J. Roy. Stat. Soc., Supp.*, 9, 128-139.
- (1950): The theory of fractional replication in factorial experiments. *Sankhyā*, 10, 81-86.
- (1951): A simplified approach to factorial experiments and the punched card technique in the construction and analysis of designs. *Bull. Int. Stat. Inst.*, 33, 1-27.
- STOLL, R. R. (1952): *Linear Algebra and Matrix Theory*, McGraw-Hill, London.
- YATES, F. (1937): The design and analysis of factorial experiments. *Tech. Commun. Imp. Bur. Soil Sci.*, No. 35.

*Paper received : February, 1958.*

*Revised : August, 1958.*

# PROPERTIES OF THE INVARIANT $I_m$ ( $m$ -odd) FOR DISTRIBUTIONS ADMITTING SUFFICIENT STATISTICS

By B. RAJA RAO

*University of Poona, India*

**SUMMARY.** In this paper, the exact form of  $I_m$  ( $m$ -odd) is obtained in general terms for the distributions admitting sufficient statistics. The case when  $m$  is odd presents some difficulties and these are got over by using a certain technique. From this, the exact form of  $I_m$  ( $m$ -odd) is deduced for the normal, the type III, the Poisson and the Binomial distributions, both when the parameters vary simultaneously and separately. Finally, it is shown that  $\sqrt{m}/I_m$ , whether  $m$  is odd or even, for the normal distribution leads to the usual prior probability forms for the parameters. Extension to multivariate distributions is immediate. Thus our results generalize those of Huzurbazar (1955).

## 1. INTRODUCTION

The statistical significance of the invariants is described by Jeffreys (1946, 1948). The invariant  $I_2$  is used by him in stating the prior probability of parameters in estimation problems and significance tests. Huzurbazar (1955) has obtained the exact form of  $I_m$  ( $m$ -even) for the distributions admitting sufficient statistics explicitly in terms of the parameters and has deduced the exact form of  $I_2$  for the type III distribution. He has also obtained (Huzurbazar, 1955) the exact form of  $I_1$  for the normal distribution  $N(\lambda, \sigma)$  when the parameters  $\lambda$  and  $\sigma$  vary separately. He has shown that  $I_1$  leads to the appropriate prior probability forms given by Jeffreys for the parameters.

## 2. DISTRIBUTIONS ADMITTING SUFFICIENT STATISTICS

We shall take the most general form of distributions admitting sufficient statistics as given by Koopman (1936) ;

$$f(x, \alpha_j) = \exp \left\{ \sum_{k=1}^p u_k(\alpha_j) v_k(x) + A(x) + B(\alpha_j) \right\} \quad \dots \quad (2.1)$$

where  $\alpha_j$  for brevity denotes the set of  $p$  parameters  $(\alpha_1, \alpha_2, \dots, \alpha_p)$ . Following Huzurbazar (1955), we have, since for all  $\alpha_j$ ,  $\int f(x, \alpha_j) dx \equiv 1$ ,

$$\int \exp \left\{ \sum_{k=1}^p u_k(\alpha_j) v_k(x) + A(x) \right\} dx = \exp \{-B(\alpha_j)\}. \quad \dots \quad (2.2)$$

Now the  $u_k(\alpha_j)$  are  $p$  independent functions of the  $p$  parameters  $\alpha_j$ . We can express the  $\alpha_j$  inversely as functions of the  $u_k$ 's. Then  $B(\alpha_j)$  can be expressed in terms of the  $u_k$ 's as  $B(\alpha_j) = b(u_k)$ , so that (2.2) may be written as

$$\int \exp \left\{ \sum_{k=1}^p u_k(\alpha_j) v_k(x) + A(x) \right\} dx = \exp \{-b(u_k)\}. \quad \dots \quad (2.3)$$



The invariant  $I_m$  is defined as  $I_m = \int |f'^{\frac{1}{m}} - f^{\frac{1}{m}}|^m dx$ , so that when  $m$  is even, it is nothing more than straightforward expansion and integration. But when  $m$  is odd, the absolute value presents some difficulties. For that we use the following technique. We split up the range of integration into two parts  $R$  and  $R^*$  throughout which the integrands are positive. For brevity, write  $f(x, \alpha_j) = f$ ,  $f(x, \alpha'_j) = f'$ ,  $u_k(\alpha'_j) = u'_k$ , etc.

Now  $f' > f$  if, and only if,

$$\sum_{k=1}^p v_k(x)(u'_k - u_k) > B - B'. \quad \dots (2.4)$$

Let  $R$  be the set of all points  $x$  in the interval  $(-\infty, \infty)$  for which the inequality (2.4) holds and let  $R^*$  be the set of the remaining points, so that  $f' \leq f$  for all  $x$  belonging to  $R^*$ . Then, since  $m$  is odd,

$$\begin{aligned} I_m &= \int_R \left( f'^{\frac{1}{m}} - f^{\frac{1}{m}} \right)^m dx - \int_{R^*} \left( f'^{\frac{1}{m}} - f^{\frac{1}{m}} \right)^m dx \\ &= \int_{-\infty}^{\infty} \left( f'^{\frac{1}{m}} - f^{\frac{1}{m}} \right)^m dx - 2 \int_{R^*} \left( f'^{\frac{1}{m}} - f^{\frac{1}{m}} \right)^m dx. \quad \dots (2.5) \end{aligned}$$

Following Huzurbazar, we shall get,

$$\begin{aligned} I_m &= \sum_{\gamma=0}^m (-)^{\gamma} \binom{m}{\gamma} \exp \left\{ \frac{m-\gamma}{m} B' + \frac{\gamma}{m} B - b \left( \frac{m-\gamma}{m} u'_k + \frac{\gamma}{m} u_k \right) \right\} - \\ &- 2 \sum_{\gamma=0}^m (-)^{\gamma} \binom{m}{\gamma} \exp \left\{ \frac{m-\gamma}{m} B' + \frac{\gamma}{m} B - b^* \left( \frac{m-\gamma}{m} u'_k + \frac{\gamma}{m} u_k \right) \right\}, \quad \dots (2.6) \end{aligned}$$

where the function  $b^*(u_k)$  is defined by [the relation analogous to (2.3)],

$$\int_{R^*} \exp \left\{ \sum_{k=1}^p u_k(\alpha_j) v_k(x) + A(x) \right\} dx = \exp \{ -b^*(u_k) \}. \quad \dots (2.7)$$

It may be pointed out here that the exact form of  $I_m$  ( $m$ -odd) cannot be expressed explicitly in terms of the parameters due to the presence of the function  $b^*$  in (2.6). But in the case of many particular distributions,  $b^*$  can actually be evaluated as an explicit function of the parameters as is shown in the examples discussed below. It is interesting to note that the first series in (2.6) is just the exact form of  $I_m$  ( $m$ -even) obtained by Huzurbazar (1955) [his equation (23)].

# THE INVARIANT FOR DISTRIBUTIONS ADMITTING SUFFICIENT STATISTICS

*Exact form of  $I_m$  ( $m$ -odd) for the normal distribution.* We shall now deduce the exact form of  $I_m$  ( $m$ -odd) for the normal distribution  $N(\lambda, \sigma)$  from (2.6) when the parameters  $\lambda$  and  $\sigma$  vary simultaneously. We have

$$f(x, \lambda, \sigma) = \exp \left\{ -\frac{x^2}{2\sigma^2} + \frac{\lambda x}{\sigma^2} - \frac{\lambda^2}{2\sigma^2} - \log \sigma \sqrt{(2\pi)} \right\}, (-\infty < x < \infty). \quad \dots (2.8)$$

Here  $u_1 = \frac{1}{\sigma^2}$ ,  $u_2 = \frac{\lambda}{\sigma^2}$  and  $B(\lambda, \sigma) = - \left( -\frac{\lambda^2}{2\sigma^2} + \log \sigma \sqrt{(2\pi)} \right)$

$$= - \left( \frac{u_2^2}{2u_1} + \log \sqrt{\frac{2\pi}{u_1}} \right) = b(u_1 u_2).$$

It is easy to see that  $R^*$  is given by the set of all points  $x$  for which

$$x^2(\sigma^2 - \sigma'^2) - 2x(\lambda'\sigma^2 - \lambda\sigma'^2) + \lambda'^2\sigma^2 - \lambda^2\sigma'^2 - \rho^2(\sigma^2 - \sigma'^2) \geq 0 \quad \dots (2.9)$$

$$\rho^2 = \frac{2\sigma^2\sigma'^2 \log \frac{\sigma'}{\sigma}}{\sigma'^2 - \sigma^2}. \quad \dots (2.10)$$

where

Two cases arise according as  $\sigma > \sigma'$  or  $\sigma < \sigma'$ .

*Case (1):* Let  $\sigma > \sigma'$ . Then  $R^*$  is composed of two intervals  $(-\infty, \mu_1)$  and  $(\mu_2, \infty)$  where  $\mu_1$  and  $\mu_2$  ( $\mu_1 < \mu_2$ ) are the roots of the quadratic

$$x^2(\sigma^2 - \sigma'^2) - 2x(\lambda'\sigma^2 - \lambda\sigma'^2) + \lambda'^2\sigma^2 - \lambda^2\sigma'^2 - \rho^2(\sigma^2 - \sigma'^2) = 0 \quad \dots (2.11)$$

given by

$$x = \mu_{1,2} = \frac{\lambda'\sigma^2 - \lambda\sigma'^2 \pm K}{\sigma^2 - \sigma'^2} \quad \text{where } K = +\sqrt{\rho^2(\sigma^2 - \sigma'^2)^2 + (\lambda' - \lambda)^2\sigma^2\sigma'^2}. \quad \dots (2.12)$$

Substituting in (2.6) we find after some reduction

$$I_m = \sum_{\gamma=0}^m \frac{(-1)^\gamma \binom{m}{\gamma}}{\sigma' \frac{m-\gamma}{m} \sigma \frac{\gamma}{m}} \delta_\gamma \cdot \exp \left\{ \frac{-\gamma(m-\gamma)(\lambda' - \lambda)^2 \delta_\gamma^2}{2m^2\sigma^2\sigma'^2} \right\} \left[ 2 \int_{t_1}^{t_2} \frac{1}{\sqrt{(2\pi)}} e^{-t^2/2} dt - 1 \right] \dots (2.13)$$

$$\delta_\gamma^2 = \frac{m\sigma^2\sigma'^2}{(m-\gamma)\sigma^2 + \gamma\sigma'^2} \quad \dots (2.14)$$

where

$$t_1 = \frac{(\lambda' - \lambda)\delta_\gamma - K}{(\sigma^2 - \sigma'^2)\delta_\gamma} \quad \text{and} \quad t_2 = \frac{(\lambda' - \lambda)\delta_\gamma + K}{(\sigma^2 - \sigma'^2)\delta_\gamma}. \quad \dots (2.15)$$



In terms of the error function, we can write (2.13) as

$$I_m = \sum_{\gamma=0}^m \frac{(-)^\gamma \binom{m}{\gamma}}{\sigma' \frac{m-\gamma}{m} \sigma \frac{\gamma}{m}} \cdot \delta_\gamma \cdot \exp \left\{ \frac{-\gamma(m-\gamma)(\lambda' - \lambda)^2 \delta_\gamma^2}{2m^2 \sigma^2 \sigma'^2} \right\} \left[ \operatorname{erf} \left( \frac{t_2}{\sqrt{2}} \right) - \operatorname{erf} \left( \frac{t_1}{\sqrt{2}} \right) - 1 \right] \quad \dots (2.16)$$

Case (2): Let  $\sigma < \sigma'$ . Then  $R^*$  is the interval  $(\mu_1, \mu_2)$  and in this case the exact form of  $I_m(m\text{-odd})$  would be given by just the negative of (2.13), so that in the general case, the absolute value of (2.13) gives the exact form of  $I_m(m\text{-odd})$  when both the parameters  $\lambda$  and  $\sigma$  vary simultaneously.

Putting  $m = 1$  in the above, we have immediately, since  $\delta_0 = \sigma'$  and  $\delta_1 = \sigma$ ,

$$I_1 = \left| \operatorname{erf} \left( \frac{(\lambda' - \lambda)\sigma' + K}{(\sigma^2 - \sigma'^2)\sqrt{2} \cdot \sigma'} \right) - \operatorname{erf} \left( \frac{(\lambda' - \lambda)\sigma' - K}{(\sigma^2 - \sigma'^2)\sqrt{2} \cdot \sigma'} \right) + \operatorname{erf} \left( \frac{(\lambda' - \lambda)\sigma - K}{(\sigma^2 - \sigma'^2)\sqrt{2} \cdot \sigma} \right) - \operatorname{erf} \left( \frac{(\lambda' - \lambda)\sigma + K}{(\sigma^2 - \sigma'^2)\sqrt{2} \cdot \sigma} \right) \right|. \quad \dots (2.17)$$

Exact form of  $I_m(m\text{-odd})$  when the parameters  $\lambda$  and  $\sigma$  vary separately. Let us now deduce the exact form of  $I_m(m\text{-odd})$  when the parameters  $\lambda$  and  $\sigma$  vary separately. First suppose that  $\lambda$  is kept fixed and that only  $\sigma$  varies. Then  $R^*$  is given by the set of points  $x$  for which  $|x - \lambda| \geq \rho$  when  $\sigma > \sigma'$  and  $|x - \lambda| \leq \rho$  when  $\sigma < \sigma'$ . Also, since from (2.12),  $K = \rho(\sigma^2 - \sigma'^2)$  we find

$$I_m = \left| \sum_{\gamma=1}^m \frac{(-)^\gamma \binom{m}{\gamma}}{\sigma' \frac{m-\gamma}{m} \sigma \frac{\gamma}{m}} \cdot \delta_\gamma \cdot \left[ 2 \operatorname{erf} \left( \frac{\rho}{\sqrt{2} \cdot \delta_\gamma} \right) - 1 \right] \right|. \quad \dots (2.18)$$

Next suppose that  $\sigma$  is kept fixed and that only  $\lambda$  varies. Then one of the roots of the equation (2.4) is infinite and the other is  $x = \frac{\lambda + \lambda'}{2}$ . In this case  $R^*$  is the interval  $\left(-\infty, \frac{\lambda + \lambda'}{2}\right)$  when  $\lambda' > \lambda$  and  $\left(\frac{\lambda + \lambda'}{2}, \infty\right)$  when  $\lambda' < \lambda$ . Considering separately the two cases when  $\lambda' > \lambda$  and  $\lambda' < \lambda$ , it may be seen that the exact form of  $I_m(m\text{-odd})$  is given by

$$I_m = \sum_{\gamma=0}^m (-)^\gamma \binom{m}{\gamma} \exp \left\{ \frac{-\gamma(m-\gamma)(\lambda' - \lambda)^2}{2m^2 \sigma^2} \right\} \operatorname{erf} \left( \frac{(m-2\gamma)|\lambda' - \lambda|}{2\sqrt{2} \cdot m\sigma} \right). \quad \dots (2.19)$$

Putting, in particular,  $m=1$  in (2.18) and (2.19), one obtains the results of Huzurbazar (1955).

Type III distribution. Consider the type III distribution :

$$f(x, a, p) = \exp\{-ax + p \log x - \log x + p \log a - \log \Gamma(p)\} \quad (x > 0). \quad \dots (2.20)$$

Here  $u_1 = a$  and  $u_2 = p$  and  $B(a, p) = p \log a - \log \Gamma(p) = u_2 \log u_1 - \log \Gamma(u_2)$ .

# THE INVARIANT FOR DISTRIBUTIONS ADMITTING SUFFICIENT STATISTICS

In this case  $R^*$  is given by the set of points  $x$  for which

$$e^{-(a'-a)x} \cdot x^{p'-p} - \frac{a^p \Gamma(p')}{a'^{p'} \Gamma(p)} \leq 0. \quad \dots (2.21)$$

For brevity we discuss below only the following two interesting cases ( $a' > a$ ,  $p'-p > -1$ ) and ( $a' \leq a$ ,  $p'-p \leq -1$ ). Let now  $\mu_1$  and  $\mu_2$  ( $\mu_1 < \mu_2$ ) be the roots of the equation

$$e^{-(a'-a)x} \cdot x^{p'-p} - \frac{a^p \Gamma(p')}{a'^{p'} \Gamma(p)} = 0. \quad \dots (2.22)$$

Case (1): Let  $a' > a$  and  $p'-p > -1$ . Then  $R^*$  is composed of the two intervals  $(0, \mu_1)$  and  $(\mu_2, \infty)$ .

Substituting in the relation (2.6), we find after some simplification,

$$I_m = \sum_{\gamma=0}^m (-1)^\gamma \binom{m}{\gamma} \frac{a'^{\frac{m-\gamma}{m} p'} \cdot a^{\frac{\gamma}{m} p}}{[\Gamma(p')]^{\frac{m-\gamma}{m}} [\Gamma(p)]^{\frac{\gamma}{m}}} \cdot \frac{\Gamma\left(\frac{\overline{m-\gamma p'} + \gamma p}{m}\right)}{\left(\frac{\overline{m-\gamma a'} + \gamma a}{m}\right)^{\frac{\overline{m-\gamma p'} + \gamma p}{m}}} \{2(g_{\mu_2} - g_{\mu_1}) - 1\} \quad \dots (2.23)$$

where  $g_\alpha = g_\alpha\left(\frac{\overline{m-\gamma a'} + \gamma a}{m}, \frac{\overline{m-\gamma p'} + \gamma p}{m}\right)$  is the incomplete gamma integral defined by

$$g_\alpha(a, p) = \frac{a^p}{\Gamma(p)} \int_0^a x^{p-1} e^{-ax} dx,$$

which is extensively tabulated by Pearson.

Case (2): Let  $a' \leq a$  and  $p'-p \leq -1$ . Then  $R^*$  is the interval  $(\mu_1, \mu_2)$  and in this case  $I_m$  ( $m$ -odd) would just be the negative of (2.23) so that in the general case the absolute value of (2.23) gives the exact form of  $I_m$  ( $m$ -odd) when the parameters  $a$  and  $p$  vary simultaneously. As in the previous example we may obtain the exact form of  $I_m$  ( $m$ -odd) when the parameters vary separately by putting  $a = a'$  and  $p = p'$  in (2.23) successively, and it will be seen that in this case the expressions will be greatly simplified.

*Poisson distribution.* Take the Poisson distribution, defined by

$$f(x, \alpha) = e^{-\alpha} \frac{\alpha^x}{x!} = \exp \{x \log \alpha - \log x! - \alpha\}, \quad x = 0, 1, 2, \dots \quad \dots (2.24)$$



Here  $u = \log \alpha$ ,  $b(\alpha) = -\alpha = -e^u = b(u)$  and  $R^*$  is given by the set of all points  $x$  for which  $x \leq \rho$  if  $\alpha' > \alpha$  and  $x \geq \rho$  if  $\alpha' < \alpha$  where  $\rho$  is the greatest integer contained in  $\frac{\alpha' - \alpha}{\log \frac{\alpha'}{\alpha}}$ . It will be seen that the exact form of  $I_m(m\text{-odd})$  is the absolute value of

$$\sum_{\gamma=0}^m (-)^{\gamma} \binom{m}{\gamma} \exp \left\{ \alpha'^{\frac{m-\gamma}{m}} \alpha^{\frac{\gamma}{m}} - \frac{m-\gamma}{m} \alpha' + \frac{\gamma}{m} \alpha \right\} (1-2P_{\rho}) \quad \dots (2.25)$$

where  $P_{\rho} = \text{Prob}(x \leq \rho)$ ,  $x$  being a Poisson variate with parameter  $\alpha'^{\frac{m-\gamma}{m}} \alpha^{\frac{\gamma}{m}}$ , which is extensively tabulated in the *Biometrika* tables.

It will be seen that

$$I_1 = 2 \left| \sum_{x=0}^{\rho} e^{-\alpha'} \frac{\alpha'^x}{x!} - \sum_{x=0}^{\rho} e^{-\alpha} \frac{\alpha^x}{x!} \right|$$

and that its differential form is

$$I_1 \doteq F(\alpha) d\alpha,$$

so that  $I_1$  does not lead to the usual prior probability form  $\frac{d\alpha}{\alpha}$  for  $\alpha$ , as given by Jeffreys.

*Binomial distribution.* Consider the Binomial distribution:

$$f(x, p) = \binom{n}{x} \left( \frac{p}{1-p} \right)^x (1-p)^n = \exp \left\{ x \log \frac{p}{1-p} + \log \binom{n}{x} + \log (1-p)^n \right\},$$

$$x = 0, 1, \dots, n. \quad \dots (2.26)$$

Here  $u = \log \frac{p}{1-p}$  and  $B(p) = \log (1-p)^n = -\log (1+e^u)^n = b(u)$ ,

and in this case  $R^*$  will be given by the set of all points  $x$  for which  $x \leq \delta$  if  $p' > p$  and  $x \geq \delta$  if  $p' < p$  where  $\delta$  is the greatest integer contained in  $\log \left( \frac{1-p}{1-p'} \right)^n \left[ \log \frac{p'(1-p)}{p(1-p')} \right]^{-1}$ . It is easy to see that the exact form of  $I_m(m\text{-odd})$  would be given by the absolute value of

$$\sum_{\gamma=0}^m (-)^{\gamma} \binom{m}{\gamma} \left[ (1-p')^{\frac{m-\gamma}{m}} (1-p)^{\frac{\gamma}{m}} + p'^{\frac{m-\gamma}{m}} p^{\frac{\gamma}{m}} \right]^n (1-2P_{\delta}), \quad \dots (2.27)$$

# THE INVARIANT FOR DISTRIBUTIONS ADMITTING SUFFICIENT STATISTICS

where  $P_\delta = \text{Prob } (x \leq \delta)$ ,  $x$  being a Binomial variate with the parameter

$$p' \frac{m-\gamma}{m} p \frac{\gamma}{m} \left[ (1-p') \frac{m-\gamma}{m} (1-p) \frac{\gamma}{m} + p' \frac{m-\gamma}{m} p \frac{\gamma}{m} \right]^{-1}.$$

$P_\delta$  may be obtained readily from a table of the cumulative binomial tables (*National Bureau of Standards*), or from a table of the incomplete beta function, since,

$$\sum_{j=0}^K \binom{n}{j} x^j (1-x)^{n-j} = B_{1-x}(n-K, K+1).$$

Putting in particular,  $m = 1$  in (2.27) it is seen that

$$I_1 = 2 \left| B_{1-p}(n-\delta, \delta+1) - B_{1-p'}(n-\delta, \delta+1) \right|$$

and that the differential form for  $I_1$  is

$$I_1 \doteq G(p)dp$$

so that in this case also  $I_1$  does not lead to the appropriate uniform prior probability form for the parameter  $p$ , as given by Jeffreys

Finally we shall obtain the differential forms for the parameters  $\lambda$  and  $\sigma$  of the normal distribution  $N(\lambda, \sigma)$  and show that  $\sqrt[m]{I_m}$ , whether  $m$  is odd or even, leads to the usual prior probability forms for the parameters,  $\lambda$  and  $\sigma$ , when they vary separately. But the difficulty arises, as noticed by Jeffreys, when they vary simultaneously.

First, we shall obtain the differential form of  $I_m$  for small variations of  $\lambda$ . Let  $m$  be odd. Then to the first order, we have

$$I_m \doteq (d\lambda)^m \int_{-\infty}^{\infty} \left| \frac{\partial f^{1/m}}{\partial \lambda} \right|^m dx = 2 \frac{(d\lambda)^m}{(m\sigma)^m} \int_0^{\infty} t^m \cdot \frac{e^{-t^2/2}}{\sqrt{(2\pi)}} dt = \text{const. } (d\lambda)^m, \quad \dots (2.28)$$

so that

$$\sqrt[m]{I_m} \propto d\lambda$$

which leads to the appropriate prior probability form for  $\lambda$  when  $\sigma$  is known, as given by Jeffreys.

Similarly when  $\lambda$  is known,

$$I_m \doteq (d\sigma)^m \int_{-\infty}^{\infty} \left| \frac{\partial f^{1/m}}{\partial \sigma} \right|^m dx = \text{const. } \left( \frac{d\sigma}{\sigma} \right)^m \quad \dots (2.29)$$

so that

$$\sqrt[m]{I_m} \propto \frac{d\sigma}{\sigma}$$

which again leads to the usual prior probability form for  $\sigma$  when  $\lambda$  is known, as given by Jeffreys. These results hold good even when  $m$  is even.



In conclusion, I wish to acknowledge my indebtedness to Dr. V. S. Huzurbazar for his valuable guidance and useful suggestions, which have greatly improved both the form and content of this paper. My grateful thanks are due to the Government of India for awarding me a Senior Research Training Scholarship.

#### REFERENCES

- HUZURBAZAR, V. S. (1955) : *Biometrika*, **42**, 533—537.
- JEFFREYS, H. (1946) : *Proc. Roy. Soc. A*, **186**, 453—461.
- (1948) : *Theory of Probability*, 2nd edition, Oxford, Clarendon Press.
- (1955) : *Journal of the University of Poona*, Science Section, **6**, 115—121.
- KOOPMAN, B. O. (1936) : *Trans. Amer. Math. Soc.*, **39**, 399—409.
- National Bureau of Standards (1950) : *Tables of the Binomial Probability Distribution*, U. S. Government Printing Office, Washington, D.C.

*Paper received : August, 1957.*

# NUMERICAL EVALUATION OF CERTAIN MULTIVARIATE NORMAL INTEGRALS

By PETER IHM

*Botanic Institute, Freiburg, Germany*

**SUMMARY.** A numerical solution of multivariate normal integrals over a region  $B$  is given where the covariance matrix is the sum of a diagonal matrix  $\mathbf{D}$  and the product of a row vector with its transpose. An  $n$ -fold integral over  $B$  of a multivariate normal distribution with covariance matrix  $\mathbf{D}$  is calculated in dependence of a parameter  $\tau$ , the resulting function multiplied with a function of  $\tau$  and integrated with respect to  $\tau$ . The method covers the integral of all multivariate normal distributions with density constant over the surface of a hyperellipsoid rotationally symmetric about the longest axis.

The evaluation of the multivariate normal integral has been the object of repeated studies (see David (1953), Plackett (1954), McFadden (1956)). A reduction formula given by Plackett is, theoretically, applicable to all  $n$ -dimensional normal distributions, but is quite laborious for higher  $n$ . We cannot, therefore, apply this formula conveniently for  $n$  larger than five. Another general procedure may be constructed by use of an  $n$ -dimensional extension of Simpson's rule by Von Mises (1954); but in this case also the effort in calculations is too great to make it useful for higher dimensions. For the moment the most satisfying general method seems to be the Monte Carlo method by use of an electronic computer since the number of points to be tested is independent of the dimensionality  $n$ . The author obtained good results by aid of the IBM Magnetic Drum Calculator Type 650 by generation of  $n$ -dimensional vectors with normally distributed components, but even for such a machine the amount of time necessary to reach a desired accuracy may be too great for higher  $n$ . It seems therefore justifiable to seek for simpler methods which apply at least for more special types of frequently occurring normal distributions. In this paper a simple method will be given for the integration of an  $n$ -dimensional normal distribution, the covariance matrix of which is of the form

$$\mathbf{A} = \mathbf{D} + \frac{1}{c^2} \mathbf{i} \mathbf{i}'$$

where  $\mathbf{D}$  is a positive definite diagonal matrix,  $\mathbf{i}$  a unit vector and  $c^2 > 0$  a scalar.

Let us consider the  $(n+1)$ -fold integral

$$I = \int_{-\infty}^{\infty} \frac{c}{\sqrt{2\pi}} e^{-\frac{1}{2}c^2\tau^2} \int_B \frac{1}{(2\pi)^{n/2} |\mathbf{D}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{z}-\tau\mathbf{i})' \mathbf{D}^{-1}(\mathbf{z}-\tau\mathbf{i})} dz_1 \dots dz_n d\tau \quad \dots (1)$$

where the second integral sign stands for the  $n$ -fold integral over the region  $B$ . We obtain

$$\begin{aligned} I &= \frac{c}{(2\pi)^{\frac{n+1}{2}} |\mathbf{D}|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \int_B e^{-\frac{1}{2}\{(\mathbf{z}-\tau\mathbf{i})' \mathbf{D}^{-1}(\mathbf{z}-\tau\mathbf{i}) + c^2\tau^2\}} dz_1 \dots dz_n d\tau \\ &= \frac{c}{(2\pi)^{\frac{n+1}{2}} |\mathbf{D}|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \int_B e^{-\frac{1}{2}\{\mathbf{z}' \mathbf{D}^{-1} \mathbf{z} - 2\mathbf{i}' \mathbf{D}^{-1} \mathbf{z} \tau + a^2 \tau^2\}} dz_1 \dots dz_n d\tau \\ &= \frac{c}{(2\pi)^{\frac{n+1}{2}} |\mathbf{D}|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \int_B e^{-\frac{1}{2}\{\mathbf{z}' \mathbf{A}^{-1} \mathbf{z} + a^2(\tau - a^2 \mathbf{i}' \mathbf{D}^{-1} \mathbf{z})^2\}} dz_1 \dots dz_n d\tau \\ &\quad \mathbf{A}^{-1} = \mathbf{D}^{-1} - a^2 \mathbf{D}^{-1} \mathbf{i} \mathbf{i}' \mathbf{D}^{-1} \end{aligned}$$

with



and

$$a^2 = \mathbf{i}' \mathbf{D}^{-1} \mathbf{i} + c^2. \quad \dots (2)$$

We get, by some calculations,

$$\mathbf{A} = \mathbf{D} + c^{-2} \mathbf{i} \mathbf{i}' \quad \dots (3)$$

which may be proved by direct calculation. We have

$$\begin{aligned} \mathbf{A} \mathbf{A}^{-1} &= (\mathbf{D} + c^{-2} \mathbf{i} \mathbf{i}') (\mathbf{D}^{-1} - a^{-2} \mathbf{i} \mathbf{i}' \mathbf{D}^{-1}) \\ &= \mathbf{I} - a^{-2} \mathbf{i} \mathbf{i}' \mathbf{D}^{-1} + c^{-2} \mathbf{i} \mathbf{i}' \mathbf{D} - a^{-2} c^{-2} \mathbf{i} \mathbf{i}' \mathbf{D}^{-1} \mathbf{i} \mathbf{i}' \mathbf{D}^{-1} \quad \dots (4) \end{aligned}$$

where  $\mathbf{I}$  is the unit matrix. The last three terms of (4) should give the null matrix  $\mathbf{O}$ , i.e. we should have

$$-c^2 \mathbf{i} \mathbf{i}' \mathbf{D}^{-1} + a^2 \mathbf{i} \mathbf{i}' \mathbf{D}^{-1} - \mathbf{i} \mathbf{i}' \mathbf{D}^{-1} \mathbf{i} \mathbf{i}' \mathbf{D}^{-1} = \mathbf{O}.$$

Because of (2) we obtain

$$(\mathbf{i}' \mathbf{D}^{-1} \mathbf{i}) \mathbf{i} \mathbf{i}' \mathbf{D}^{-1} - \mathbf{i} \mathbf{i}' \mathbf{D}^{-1} \mathbf{i} \mathbf{i}' \mathbf{D}^{-1}$$

and after application of the commutative law for scalar matrix multiplication to the first term and of the associative law for matrix multiplication to the second finally

$$\mathbf{i} (\mathbf{i}' \mathbf{D}^{-1} \mathbf{i}) \mathbf{i}' \mathbf{D}^{-1} - \mathbf{i} (\mathbf{i}' \mathbf{D}^{-1} \mathbf{i}) \mathbf{i}' \mathbf{D}^{-1} = \mathbf{O}$$

which proves (3).

We may in (1) invert the sequence of integration and first integrate over  $\tau$ . This yields

$$I = \frac{1}{(2\pi)^{n/2} |\mathbf{A}|^{1/2}} \int_B e^{-\frac{1}{2} \mathbf{z}' \mathbf{A}^{-1} \mathbf{z}} dz_1 \dots dz_n$$

so that  $I$  is equal to an  $n$ -fold integral of a normal distribution with expectations  $E \mathbf{z} = \mathbf{0}$  and  $E \mathbf{z} \mathbf{z}' = \mathbf{A}$ ,  $\mathbf{0}$  being the null vector. Now, if  $\mathbf{D}$  is supposed to be a diagonal matrix and  $B$  an  $n$ -dimensional interval, it is easy to compute

$$I(\tau) = \frac{1}{(2\pi)^{n/2} |\mathbf{D}|^{1/2}} \int_B e^{-\frac{1}{2} (\mathbf{z} - \tau \mathbf{i})' \mathbf{D}^{-1} (\mathbf{z} - \tau \mathbf{i})} dz_1 \dots dz_n \quad \dots (5)$$

in dependence of  $\tau$ . The integral (1) is equal to

$$I = \frac{c}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2} c^2 \tau^2} I(\tau) d\tau \quad \dots (6)$$

which may be obtained by simple product integration using Simpson's rule.

A very convenient method is the use of a Stieltjes coordinate,  $x$  for the abscissa,  $\tau$  where

$$\mathbf{x} = \frac{c}{\sqrt{2\pi}} \int_{-\infty}^{\tau} e^{-\frac{1}{2} c^2 \xi^2} d\xi.$$

# NUMERICAL EVALUATION OF CERTAIN MULTIVARIATE NORMAL INTEGRALS

$I(\tau)$  is then drawn over the transformed abscissa and the integral  $I$  is obtained by planimeter. This gives a relative error of 1 to 2 per cent. It is convenient to construct the Stieltjes integral abscissa for all values of  $c$  by substituting  $\tau^*/c$  for  $\tau$ . We get from (6)

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2} \tau^{*2}} I\left(\frac{\tau^*}{c}\right) d\tau^*, \quad \dots (7)$$

$$\tau^* = c\tau.$$

It is not necessary that  $B$  be an interval nor that  $\mathbf{D}$  be a diagonal matrix, but the method is more easily applicable in cases where it is.  $B$  and the positive definite matrix  $\mathbf{D}$  may be of either form. It is only necessary that  $c^2 > 0$ .  $c^2 = 0$  is trivial and  $c^2 < 0$  changes the positive definiteness of the quadratic form in the exponent of the normal distribution in (1) so that the integral (1) does no longer exist.

If, for  $b^2 > 0$ ,

$$\mathbf{A} = b^2 \mathbf{I} + \frac{1}{c^2} \mathbf{ii}'$$

we have the class of multivariate normal distributions with density constant over the surface of a hyperellipsoid which is rotationally symmetric about the longest axis. This may be shown by introducing

$$\mathbf{y} = \begin{pmatrix} \mathbf{i}' \\ \mathbf{R} \end{pmatrix} \mathbf{z} = \mathbf{S}\mathbf{z}, \text{ say,}$$

where  $\mathbf{S}$  is an orthogonal matrix. The covariance matrix of  $\mathbf{y}$  is

$$\mathbf{S} \left( b^2 \mathbf{I} + \frac{1}{c^2} \mathbf{ii}' \right) \mathbf{S}' = b^2 \mathbf{I} + \frac{1}{c^2} \begin{pmatrix} 1 & \mathbf{o}' \\ \mathbf{o} & \mathbf{0} \end{pmatrix}.$$

Thus, the length of one axis is proportional to  $\sqrt{(b^2 + c^{-2})}$ , the length of the others to  $b$ .

As an example consider the integral over a distribution with  $c^2 = 1/2$  and

$$\mathbf{D} = \mathbf{I}, \mathbf{i} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + 2 \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix},$$

where  $B$  is defined by  $z_1, z_2 \geq 0$ . By theory the integral must be

$$I = \frac{1}{4} + \frac{1}{2\pi} \arcsin \rho = \frac{1}{4} + \frac{1}{2\pi} \arcsin \frac{1}{2} = \frac{1}{3}$$



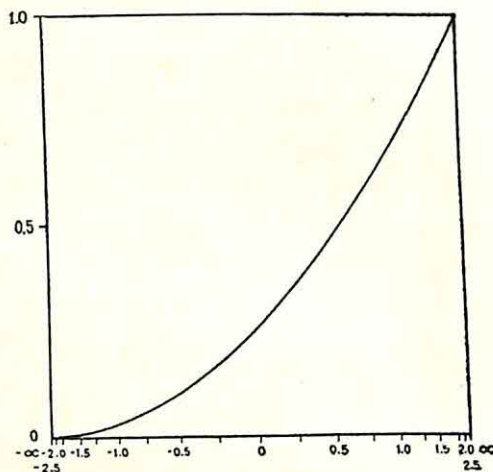
(see Cramér 1946, p. 290);  $\rho$  is the correlation coefficient. Since  $\mathbf{D} = \mathbf{D}^{-1} = \mathbf{I}$  we get from (6)

$$I(\tau) = \left\{ \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}(z-\tau/\sqrt{2})^2} dz \right\}^2 = \phi^2 \left( \frac{\tau}{\sqrt{2}} \right)$$

or for use of (7)

$$I(\tau^* \sqrt{2}) = \phi^2(\tau^*),$$

where  $\phi(x)$  denotes the normal distribution function.



The figure shows the values of  $\phi^2(\tau^*)$  in dependence of  $\tau$  drawn on Stieltjes integral coordinates. Planimeter reading gives  $I = 0.334$  which is in good agreement with expectation.

#### REFERENCES

- CRAMÉR, H. (1946): *Mathematical Methods of Statistics*, Princeton University Press.
- DAVID, F. N. (1956): A note on the evaluation of the multivariate normal integral. *Biometrika* **43**, 206-207.
- McFADDEN, J. A. (1956): An approximation for the symmetric quadrivariate normal integral. *Biometrika*, **43**, 206-207.
- PLACKETT, R. L. (1954): A reduction formula for normal multivariate integrals. *Biometrika*, **41**, 351-360.
- VON MISES, R. (1954): Numerische Berechnung mehrdimensionaler Integrale. *Zeitschrift für angewandte Mathematik und Mechanik*, **34**, 201-210.

*Paper received : February, 1958.*

# ON THE EVALUATION OF THE PROBABILITY INTEGRAL OF A MULTIVARIATE NORMAL DISTRIBUTION

By S. JOHN

*Indian Statistical Institute, Calcutta*

**SUMMARY.** A simple reduction formula is obtained for the probability integral of a multivariate normal distribution. The derivation involves only elementary results in probability theory. The formula obtained can be used in evaluating probability integrals of multivariate normal distributions of order  $k$  when those of order  $k-1$  are readily available. An example is worked out illustrating the use of the formula.

## INTRODUCTION

In so far as statisticians often assume many populations to be multivariate normal, the evaluation of the probability integral of the multivariate normal distribution is of especial importance. While this has been done for univariate and bivariate normal distributions, the extension to populations of higher orders presents considerable difficulties. For previous work on this problem reference may be made to David (1953), Plackett (1954), Moran (1956) and Das (1956).

## NOTATION AND PRELIMINARIES

The vector valued random variable  $X = (X_1, X_2, \dots, X_p)$  will be said to have the density function  $n_x(\mu; \Sigma)$  if it has the multivariate normal distribution with means  $\mu = (\mu_1, \mu_2, \dots, \mu_p)$  and dispersion matrix  $\Sigma \equiv (\sigma_{ij})$

$$n_x(\mu; \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right\}.$$

$\Sigma_{.i}$  will denote the matrix  $(\sigma_{rs} - \sigma_{ir}\sigma_{is}/\sigma_{ii})$  with the  $i$ -th row and column deleted. For scalar  $u$ ,

$$\mu_{.i}(u) = (\mu_1, \mu_2, \dots, \mu_{i-1}, \mu_{i+1}, \dots, \mu_p) + (u - \mu_i)(\beta_{1i}, \dots, \beta_{i-1,i}, \beta_{i+1,i}, \dots, \beta_{pi})$$

where  $\beta_{rs} = \sigma_{rs}/\sigma_{ss}$ .

Given  $X_i = u$ ,  $X_{.i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p)$  is distributed according to the density function  $n_{x_{.i}}(\mu_{.i}(u), \Sigma_{.i})$ .

## THE NEW METHOD

The problem is to evaluate integrals of the form

$$\int_{a_1}^{\infty} \int_{a_2}^{\infty} \dots \int_{a_p}^{\infty} n_x(\mu; \Sigma) dx \quad \text{and} \quad \int_{-\infty}^{b_1} \int_{-\infty}^{b_2} \dots \int_{-\infty}^{b_p} n_x(\mu; \Sigma) dx \quad \dots (1)$$

$$\int_{a_1}^{\infty} \dots \int_{a_p}^{\infty} n_x(\mu; \Sigma) dx = \int_0^{\infty} \dots \int_0^{\infty} n_x(\mu - a; \Sigma) dx \quad \dots (2)$$

$$\int_{-\infty}^{b_1} \dots \int_{-\infty}^{b_p} n_x(\mu; \Sigma) dx = \int_0^{\infty} \dots \int_0^{\infty} n_x(b - \mu; \Sigma) dx \quad \dots (3)$$

and,

$$a = (a_1, \dots, a_p) \quad \text{and} \quad b = (b_1, \dots, b_p). \quad \dots (4)$$

where



Therefore we need consider only integrals of the type

$$I = \int_0^\infty \dots \int_0^\infty n_x(\mu; \Sigma) dx. \quad \dots (5)$$

We now observe that  $I$  is the probability that  $U \equiv \min(X_1, \dots, X_p)$  is greater than or equal to zero. If  $f(u)$  is the probability density function of  $U$ ,

$$I = \int_0^\infty f(u) du. \quad \dots (6)$$

To derive the probability density function of  $U$ , we employ the following simple argument. The event that  $U$  lies in the interval  $(u-du, u)$  can happen in  $p$  mutually exclusive ways. Either  $X_1$  lies in the interval  $(u-du, u)$  and  $X_2, X_3, X_4, \dots, X_p$  are all greater than  $u$  or  $X_2$  lies in the interval  $(u-du, u)$  and  $X_1, X_3, X_4, \dots, X_p$  are all greater than  $u$  or  $X_3$  lies in the interval  $(u-du, u)$  and  $X_1, X_2, X_4, X_5, \dots, X_p$  are all greater than  $u$  and so on. Thus,

$$f(u) = \sum_{i=1}^p \left\{ \int_u^\infty \dots \int_u^\infty n_{x_i}(\mu_{.i}(u); \Sigma_{.i}) dx_{.i} \right\} (2\pi\sigma_{ii})^{-1/2} \exp \{-(u-\mu)^2/2\sigma_{ii}\} \quad \dots (7)$$

where

$$dx_{.i} = dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_p.$$

If all the variances  $\sigma_{ii}$ 's are equal and all the covariances  $\sigma_{ij}$ 's are equal and also  $\mu_1 = \mu_2 = \dots = \mu_p$  (7) will take the simpler form

$$f(u) = p(2\pi\sigma_{11})^{-1/2} \exp \{-(u-\mu_1)^2/2\sigma_{11}\} \int_u^\infty \dots \int_u^\infty n_{x_1}(\mu_{.1}(u), \Sigma_{.1}) dx_{.1} \quad \dots (9)$$

There are also obvious simplifications when some of the simple or partial correlations are zero.

The evaluation of the density function (7) requires only a knowledge of the probability integral of multivariate normal distributions of order  $p-1$ . Thus (6) may be regarded as a sort of reduction formula. When probability integrals of normal distributions of order  $k$  are readily available, formula (6) may be used in conjunction with methods of numerical integration to evaluate probability integrals of order  $k+1$  and, with more labour, of order  $k+2$ . Thus with the tables now available, probabilities for distributions of order three and four may be evaluated without much difficulty. We also feel that (6) may be profitably used to extend existing tables of the probability integral of the bivariate normal distribution (Pearson, 1931). This will require only the ordinates of the standard univariate normal curve at selected points and the area under this curve below these ordinates. Extensive tables for these are available in several places (Pearson and Hartley, 1954; U. S. Dept. of Commerce, 1953).

# PROBABILITY INTEGRAL OF A MULTIVARIATE NORMAL DISTRIBUTION

The example given below will help to clarify some points.

Example : We shall evaluate

$$I = \int_0^{\infty} \int_0^{\infty} \int_0^{\infty} n_x(\mu; \Sigma) dx_1 dx_2 dx_3 \quad \dots \quad (8)$$

for the case  $\mu = 0$ ,  $\sigma_{11} = \sigma_{22} = \sigma_{33} = 1$ ,  $\sigma_{12} = \sigma_{21} = \sigma_{13} = \sigma_{31} = \sigma_{23} = \sigma_{32} = 0.6$   
(We wish to emphasize here that our method is applicable whatever  $\mu$  and  $\Sigma$ ). Formula (6) in this case becomes

$$I = 3 \int_0^{\infty} f_1(u) f_2(u) du \quad \dots \quad (9)$$

where

$$f_1(u) = \int_{u/2}^{\infty} \int_{u/2}^{\infty} \frac{1}{2\pi[1-(.375)^2]^{\frac{1}{2}}} \exp \left[ -\frac{1}{2(1-(.375)^2)} (x_1^2 - 2 \times .375 x_1 x_2 + x_2^2) \right] dx_1 dx_2 \quad \dots \quad (10)$$

and

$$f_2(u) = (2\pi)^{-\frac{1}{2}} e^{-u^2/2}. \quad \dots \quad (11)$$

$f(u) = f_1(u) \cdot f_2(u)$  was calculated for  $u = 0, .2, .4, \dots, 4.2$  from values of  $f_1(u)$  taken from Karl Pearson's *Tables for Statisticians and Biometricians*, Part II, and values of  $f_2(u)$  taken from *Biometrika Tables*. We used Weddle's rule to calculate the integral of  $f(u)$  from 0 to 3.6 and Simpson's three-eighths rule to evaluate the integral from 3.6 to 4.2. Of course any other convenient method of numerical integration could have been adopted.

The contribution to (9) by the integral of  $f(u)$  from 4.2 to  $\infty$  remains to be assessed. From tables we find that  $f_1(u) < .0019$  for  $u \geq 4.2$ . Therefore,

$$\int_{4.2}^{\infty} f(u) du < .0019 \int_{4.2}^{\infty} (2\pi)^{-\frac{1}{2}} e^{-u^2/2} du \approx .0019 \times .0000133 \approx (.25)10^{-7} \quad \dots \quad (12)$$

This we regard negligible. Thus the value of the integral from 0 to 4.2 may be taken as an approximation to the value of the integral from zero to infinity. In general, we can always replace the upper limit of integration in (6) by a suitable finite number  $h$  which will give results of sufficient accuracy. For the problem of our example, the result obtained by this method was

$$I = .27907 \quad \dots \quad (13)$$



We now compare this result with the one given by the formula

$$(2\pi)^{-3/2} |M|^{-1/2} \int_0^\infty \int_0^\infty \int_0^\infty \exp \left\{ -\frac{1}{2} x M^{-1} x' \right\} dx = \frac{\cos^{-1}(-\rho_{12}) + \cos^{-1}(-\rho_{13}) + \cos^{-1}(-\rho_{23}) - \pi}{4\pi} \dots (14)$$

where  $M$  is the matrix  $(\rho_{ij})$  ( $\rho_{11} = \rho_{22} = \rho_{33} = 1$ ) Placket (1954). Formula (14) gives the value .27554 which differs from result (13) by about .0035. This much of difference was expected since, in evaluating  $f(u)$  for various values of  $u$ , only linear interpolation was used with regard to the correlation coefficient.

#### REFERENCES

- DAS, S. C. (1956): The numerical evaluation of a class of integrals. *11. Proc. Camb. Phil. Soc.*, 52, 442-448.
- DAVID, F. N. (1953): A note on the evaluation of the multivariate normal integral. *Biometrika*, 40, 458-459.
- MORAN, P. A. P. (1956): The numerical evaluation of a class of integrals. *Proc. Camb. Phil. Soc.*, 52, 230-233.
- PEARSON, E. S. AND HARTLEY, H. O. (1954): *Biometrika Tables for Statisticians*, Cambridge, Biometrika Trustees. Table I, 1.
- PEARSON, KARL (1931): *Tables for Statisticians and Biometricians*, Part 2, Tables VIII & IX, 78-137. Cambridge University Press.
- PLACKET, R. L. (1954): A reduction formula for normal multivariate integrals. *Biometrika*, 41, 351-360.
- U. S. Department of Commerce (1953): *Tables of Normal Probability functions*,

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \text{ and } \frac{1}{\sqrt{2\pi}} \int_{-x}^x e^{-a^2/2} da$$

*Paper received : January, 1959.*

# THE DISTRIBUTION OF WALD'S CLASSIFICATION STATISTIC WHEN THE DISPERSION MATRIX IS KNOWN

By S. JOHN

*Indian Statistical Institute, Calcutta*

**SUMMARY.** The discriminant function can be used for classifying an individual as belonging to one or other of two populations provided we know the parameters characterising the two populations. Wald suggested the use of a certain statistic in situations where such knowledge is absent. The exact distribution of this statistic in the case where the dispersion matrix is known, is obtained in this paper.

## 1. INTRODUCTION

Consider the problem of classifying a given collection of individuals as belonging to one or other of two populations  $P_1, P_2$  based on measurements carried out on each individual with respect to  $p$  characteristics  $x_1, x_2, \dots, x_p$ . Let us assume that among the individuals belonging to  $P_1$ ,  $x_1, x_2, \dots, x_p$  follow a multivariate normal distribution with means  $\mu_1, \mu_2, \dots, \mu_p$  and variance-covariance matrix  $\Sigma$  and that among individuals belonging to  $P_2$ ,  $x_1, x_2, \dots, x_p$  follow a multivariate normal distribution with the same variance-covariance matrix  $\Sigma$  but with means  $\nu_1, \nu_2, \dots, \nu_p$ . According to a method originally suggested by Fisher (1936) an individual with measurements  $y_1, y_2, \dots, y_p$  is assigned to  $P_1$  or  $P_2$  according as

$$(\mu - \nu)\Sigma^{-1}y' \begin{cases} > \\ < \end{cases} \frac{1}{2}(\mu - \nu)\Sigma^{-1}(\mu + \nu)'$$

where  $\mu = (\mu_1, \dots, \mu_p)$  and  $\nu = (\nu_1, \nu_2, \dots, \nu_p)$ . It will be noted that Fisher's method requires a knowledge of  $\mu, \nu$  and  $\Sigma$ . To remedy this Wald (1944) considered the use of the statistic

$$u = (\bar{x}^{(1)} - \bar{x}^{(2)})S^{-1}y',$$

the individuals being classified as belonging to  $P_1$  when  $U < d$ , where  $d$  is so determined that the critical region is of the desired size;  $\bar{x}^{(1)}$  is the vector of means determined from measurements on a sample of  $n_1$  individuals known to belong to  $P_1$ ;  $\bar{x}^{(2)}$  is the vector of means determined from measurements on a random sample of  $n_2$  individuals known to belong to  $P_2$ , and  $S$  is the variance-covariance matrix estimated from the pooled corrected sums of squares and products from the two samples. The distribution of  $U$  turns out to be complicated and Wald has not given an explicit expression for it.

This paper considers the distribution of  $U$  in the case when the variance-covariance matrix is known; i.e. the distribution of

$$V = (\bar{x}^{(1)} - \bar{x}^{(2)})\Sigma^{-1}y' \quad \dots \quad (1.1)$$

where  $\bar{x}^{(1)}, \bar{x}^{(2)}, \Sigma$  and  $y$  have the same meanings as before.



## 2. REDUCTION OF THE PROBLEM

Looking at the problem from a more general point of view, it will be seen that the statistic whose distribution is in question is

$$z = t \Sigma^{-1} w' \quad \dots (2.1)$$

where  $t = (t_1, t_2, \dots, t_p)$  is a vector of  $p$  normal variates following a multivariate normal distribution with variance-covariance matrix  $\Sigma$  and means  $a = (a_1, a_2, \dots, a_p)$ , and where  $w = (w_1, w_2, \dots, w_p)$  is a vector of  $p$  normal variates independent of  $t$  and following a multivariate normal distribution with variance-covariance matrix  $\Sigma$  and mean  $b = (b_1, b_2, \dots, b_p)$ . The statistic  $V$  reduces to  $z$  when multiplied by  $(n_1 n_2 / n_1 + n_2)^{\frac{1}{2}}$ .

In the further reduction of the problem we require the following.

*Lemma: The statistic  $z$  is invariant under non-singular transformations of  $t$  and  $w$  provided the two transformations are the same.*

*Proof:* Let  $x = t C$  and  $y = w C$  where  $C$  is a  $(p \times p)$  non-singular matrix. Let  $\Sigma_0$  denote the variance-covariance matrix of  $x$  (which is the same as the variance-covariance matrix of  $y$ ). Then

$$\Sigma_0 = C' \Sigma C \text{ and } x \Sigma_0^{-1} y' = t C (C^{-1} \Sigma^{-1} C'^{-1}) C' w' = t \Sigma^{-1} w'.$$

This proves the lemma.

Since, when  $\Sigma$  is positive definite, there always exists a non-singular matrix  $C$  such that

$$C' \Sigma C = 1$$

this lemma reduces our problem to a consideration of the statistic

$$T = u_1 v_1 + u_2 v_2 + \dots + u_p v_p \quad \dots (2.2)$$

where  $u_1, v_1, u_2, v_2, \dots, u_p, v_p$  are independent normal variates with unit variance. Without loss of generality we may assume that

$$E(u_1) = E(u_2) = \dots = E(u_p) = 0. \quad \dots (2.3)$$

Let

$$E(v_i) = m_i (i = 1, 2, \dots, p). \quad \dots (2.4)$$

The expectation and variance of  $T$  are easily found.

$$E(T) = \sum_{i=1}^p E(u_i) E(v_i) = 0$$

since, by our assumption

$$E(u_i) = 0 \quad (i = 1, 2, \dots, p)$$

$$V(T) = \sum_{i=1}^p V(u_i v_i) = \sum_{i=1}^p E(u_i^2 v_i^2) = \sum_{i=1}^p (1 + m_i^2) = p + \sum_{i=1}^p m_i^2.$$

To find the distribution of  $T$  we adopt the method of characteristic functions.

# WALD'S CLASSIFICATION STATISTIC FOR KNOWN DISPERSION MATRIX

## 3. THE CHARACTERISTIC FUNCTION OF $T$

The characteristic function of  $u_i v_i$  is

$$\begin{aligned}\varphi_{u_i v_i} &= E(e^{i\theta u_i v_i}) \\ &= E(e^{-\frac{1}{2}v_i^2 \theta^2}) \\ &= \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}v_i^2 \theta^2 - \frac{1}{2}(v_i - m_i)^2} dv_i \\ &= (1 + \theta^2)^{-\frac{1}{2}} \exp \left\{ -\frac{m_i^2 \theta^2}{2(1 + \theta^2)} \right\}. \quad \dots (3.1)\end{aligned}$$

and hence we get for the characteristic function of  $T$

$$\varphi(\theta) = \varphi_T(\theta) = (1 + \theta^2)^{-p/2} \exp \left\{ -\frac{m\theta^2}{1 + \theta^2} \right\}$$

where  $m = \frac{1}{2} \sum_{i=1}^p m_i^2$ .

If we denote by  $p(T)$  the density function of  $T$ , we then have, by inversion,

$$p(T) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-i\theta T} \varphi(\theta) d\theta. \quad \dots (3.2)$$

For the purpose of inversion we distinguish two cases viz (1)  $p$  even and (2)  $p$  odd.

*Inversion : Case (i),  $p = 2n$ .*

In this case,

$$\begin{aligned}\varphi(\theta) &= (1 + \theta^2)^{-n} \exp \left\{ -\frac{m\theta^2}{1 + \theta^2} \right\} \\ &= e^{-m} \left[ \frac{1}{(1 + \theta^2)^n} + \frac{1}{1!} \frac{m}{(1 + \theta^2)^{n+1}} + \frac{1}{2!} \frac{m^2}{(1 + \theta^2)^{n+2}} + \dots \right]. \quad \dots (3.3)\end{aligned}$$

Therefore,

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-i\theta T} \varphi(\theta) d\theta = \frac{e^{-m}}{2\pi} \int_{-\infty}^{+\infty} \left[ \frac{e^{-i\theta T}}{(1 + \theta^2)^n} + \frac{m}{1!} \frac{e^{-i\theta T}}{(1 + \theta^2)^{n+1}} + \dots \right] d\theta. \quad \dots (3.4)$$



The series occurring as integrand in (3.4) being uniformly convergent, permits term by term integration and we may write,

$$p(T) = e^{-m} \sum_{r=0}^{\infty} \frac{m^r}{r!} g_{n+r}(T) \quad \dots \quad (3.5)$$

where

$$\begin{aligned} g_r(T) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{e^{-i\theta T}}{(1+\theta^2)^r} d\theta \\ &= \frac{1}{2^{2r-1}(r-1)!} \cdot e^{-|T|} \left\{ (2|T|)^{r-1} + \frac{1}{1!} r(r-1)(2|T|)^{r-2} + \frac{1}{2!} (r+1)r(r-1)(r-2)(2|T|)^{r-3} + \right. \\ &\quad \left. + \dots + \frac{(2r-2)!}{(r-1)!} \right\}. \quad \dots \quad (3.6) \end{aligned}$$

(The last step is obtained by contour integration).

Case (ii),  $p = 2n+1$ .

The characteristic function of  $Z = xy$  where  $x$  and  $y$  are independent normal variates with unit variance and means zero and  $b$  respectively would be

$$\varphi_{xy}(\theta) = (1+\theta^2)^{-1/2} \exp \left\{ -a \frac{\theta^2}{1+\theta^2} \right\} \quad \dots \quad (3.7)$$

where  $a = \frac{1}{2}b^2$ . Also, the density function of  $Z$  is

$$\frac{e^{-a}}{\pi} \left[ K_0 + \frac{2|z|}{2!} K_1 a + \frac{(2z)^2}{4!} K_2 a^2 + \frac{(2|z|)^3}{6!} K_3 a^3 + \dots \right] \quad \dots \quad (3.8)$$

where  $K_\nu = K_\nu(z) = \frac{1}{2} \left( \frac{z}{2} \right)^\nu \int_0^\infty e^{-t - \frac{z^2}{4t}} \frac{dt}{t^{\nu+1}}$  [see Craig (1936)]. ... (3.9)

Using the inversion theorem for characteristic functions we get from (3.7) and (3.8)

$$\frac{e^{-a}}{2\pi} \int_{-\infty}^{+\infty} e^{-i\theta z} \frac{e^{a/(1+\theta^2)}}{(1+\theta^2)^{\frac{1}{2}}} d\theta = \frac{e^{-a}}{\pi} \left[ K_0 + \frac{K_1}{1!} |2z| a + \frac{K_2}{4!} (2z)^2 a^2 + \dots \right].$$

Therefore,

$$\int_{-\infty}^{+\infty} e^{-i\theta z} \frac{e^{a/(1+\theta^2)}}{(1+\theta^2)^{\frac{1}{2}}} d\theta = 2 \left[ K_0 + \frac{K_1}{2!} |2z| a + \frac{K_2}{4!} (2z)^2 a^2 + \dots \right] \dots \quad (3.10)$$

$$= \psi(a, z) \quad (\text{say}).$$

Differentiating both sides of (3.10)  $n$  times with respect to  $a$ , we get

$$\int_{-\infty}^{+\infty} e^{-i\theta z} \frac{e^{a/(1+\theta^2)}}{(1+\theta^2)^{\frac{1}{2}}} d\theta = \frac{\partial^n}{\partial a^n} \psi(a, z). \quad \dots \quad (3.11)$$

[We note that differentiation within the integral sign in (3.11) is permissible since

$$\int_{-\infty}^{+\infty} e^{-i\theta z} \frac{e^{-a/(1+\theta^2)}}{(1+\theta^2)^{n+\frac{1}{2}}} d\theta$$

is uniformly convergent]. Since the characteristic function of  $T$  is

$$(1+\theta^2)^{-(n+\frac{1}{2})} \exp \left\{ -\frac{m\theta^2}{1+\theta^2} \right\} = e^{-m} (1+\theta^2)^{-(n+\frac{1}{2})} \exp \{m(1+\theta^2)^{-1}\}$$

it follows from (3.11) that the density function of  $T$  is

$$p(T) = \frac{e^{-m}}{2\pi} \left[ \frac{\partial^n}{\partial a^n} \psi(a, T) \right]_{a=m}. \quad \dots \quad (3.12)$$

*Note :* The derivatives of  $\psi(a, T)$  required in (3.12) can be obtained from the relation

$$\psi(a, T) = 2 \left[ K_0 + \frac{2|T|}{2!} K_1 a + \frac{(2T)^2}{4!} K_2 a^2 + \dots \right]. \quad \dots \quad (3.13)$$

For any given value of  $T$ , the series in (3.13) can be regarded as a power series in ' $a$ '. The series is easily seen to be convergent for all values of ' $a$ ' from the ratio  $\rho_v$  of the  $(v+1)$ -th term to the  $v$ -th term

$$\rho_v = \frac{|T|}{v(2v-1)} \frac{K_v}{K_{v-1}} a = \frac{c_v a}{2v-1} \quad (\text{say}).$$

Craig in his paper referred to above, proves that  $|c_v| < 3$  for all sufficiently large values of  $v$ . Therefore  $\rho_v \rightarrow 0$  as  $v \rightarrow \infty$ . Thus the power series in (3.13) is convergent for all values



of 'a'. Therefore the derivatives of  $\psi(a, T)$  w.r.t.  $a$  can be obtained by term by term differentiation of

$$2 \left[ K_0 + \frac{2|T|}{2!} K_1 a + \frac{(2T)^2}{4!} K_2 a^2 + \dots \right].$$

Tables of  $K_\nu(T)$  can be found in Watson's book 'Theory of Bessel Functions'.

2. Some approximations to the distribution of  $T$  has been considered. Details will be given elsewhere.

#### REFERENCES

- CRAIG, C. C. (1936): On the frequency function of  $xy$ . *Ann. Math. Stat.*, **7**, 1.  
 FISHER, R. A. (1936): The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179.  
 WALD, A. (1944): On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Stat.*, **15**, 145.  
 WATSON, G. N. (1944): *Theory of Bessel Function*, Cambridge University Press.

*Paper received : October, 1958.*

*Revised : April, 1959.*

# AN EXTENSION OF HALD'S TABLE FOR THE ONE-SIDED CENSORED NORMAL DISTRIBUTION

By NIKHILESH BHATTACHARYA

*Indian Statistical Institute, Calcutta*

**SUMMARY.** Hald (1949) outlined a very convenient method of maximum likelihood estimation of the parameters of a one-sided censored normal distribution and gave tables for facilitating the process. Table 1 below is an extension of Hald's main table (Table III). Hald's table gave the values of a certain function  $z = f(h, y)$ , for values of  $h = 0.05, 0.10, \dots, 0.80$ , and for some appropriate values of  $y$ ,  $h$  being the fraction of censored observations in the sample. The present extension gives the values of  $z$  for some values of  $h$  below 0.05, for use in situations where the censored observations cannot be ignored for purposes of estimation, even though they form less than 5% of the total sample.

Hald's method of estimation is briefly as follows :

Suppose there are  $n$  observations from a normal distribution with mean  $\xi$  and variance  $\sigma^2$ , and it is known that a number, say  $a$  of these observations are less than or equal to a known point of truncation. The values of these  $a$  observations are not further specified, unlike the values above the truncation point, which may be denoted by  $x_1, x_2, \dots, x_{n-a}$ .

The point of truncation is taken as the origin. Let now

$$\xi = -\frac{\zeta}{\sigma},$$

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \quad \Phi(u) = \int_{-\infty}^u \phi(x)dx, \quad \psi(u) = \log_e \Phi(u),$$

and  $\psi'(u)$  the first derivative of  $\psi(u)$ .

Let  $h = \frac{a}{n}$  denote the observed degree of truncation in the sample.

Hald defines

$$g(h, z) = \frac{1}{\frac{h}{1-h} \psi'(z) - z},$$

$$F(h, z) = \frac{1}{2} g(h, z) [g(h, z) - z].$$

and



Let now the inverse function to  $y = F(h, z)$  with respect to  $z$  be denoted by  $z = f(h, y)$ . This function was tabulated in Table III of Hald (1949) for  $h = 0.05, 0.10, \dots, 0.80$  and for some appropriate values of  $y$ .

The estimate  $\zeta$  of  $\hat{\zeta}$  is then obtained by calculating

$$y = \frac{(n-a) \sum_{i=1}^{n-a} x_i^2}{2 \left( \sum_{i=1}^{n-a} x_i \right)^2}$$

and reading  $\hat{\zeta} = f(h, y)$  from the Table.

The next step is to calculate

$$\hat{\sigma} = g(h, \hat{\zeta}) \frac{\sum_{i=1}^{n-a} x_i}{n-a}$$

and finally

$$\hat{\xi} = -\hat{\zeta} \hat{\sigma}.$$

The function  $g(h, z)$  can be easily calculated. Table IV of Hald's paper may be used for this purpose, but direct calculation is not difficult.

The need of the present extension was felt in certain cases of fitting one-sided censored normal distributions to grouped data. The values of  $h = \frac{a}{n}$  were found to be often below 0.05, and sometimes of the order of 0.01. Although the censored part could be ignored without much loss of information, it would be desirable to make use of it, especially because for examining goodness of fit the tails are valuable. Values of  $z = f(h, y)$  in Hald's table change sharply with  $h$ , as  $h$  approaches small values. Graphical extrapolation was out of question.

The present extension intends to facilitate interpolation for values of  $h$  below 0.05. The column for  $h = 0.001$  is particularly in point. This value of  $h$  is clearly outside the range of practical interest. However, cases with  $h = 0.005$  or  $0.008$  are not uncommon and the column for  $h = 0.001$  will enable one to interpolate for such values.

The calculations were based on the Table of Normal Probability Functions, published by the National Bureau of Standards. The figures tabulated are correct to the third place of decimals.

## TABLE FOR THE ONE-SIDED CENSORED NORMAL DISTRIBUTION

The author is grateful to Shri Rabindranath Mukherjee, Shri Ramdulal Chatterjee and Shri Amal Kumar Sengupta, for the computation of the table.

### REFERENCES

HALD, A. (1949): Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Skand. Aktuar*, 119-134.

*Tables of Normal Probability Functions*, National Bureau of Standards, Applied Mathematics Series, 23.

*Paper received : August, 1958.*



TABLE 1. VALUES OF FUNCTION  $z = f(h, y)$  FOR FITTING ONE-SIDED CENSORED NORMAL DISTRIBUTIONS

$y$	$h$				
	0.001	0.010	0.020	0.035	0.050
(1)	(2)	(3)	(4)	(5)	(6)
0.500					-4.135
0.505				-4.465	-3.774
0.510				-4.039	-3.494
0.515			-4.337	-3.715	-3.268
0.520			-3.959	-3.458	-3.080
0.525	-4.421	-4.021	-3.665	-3.248	
0.530	-4.043	-3.723	-3.429	-3.072	
0.535	-3.747	-3.482	-3.232	-2.922	
0.540	-3.508	-3.283	-3.066	-2.792	
0.545	-3.310	-3.114	-2.923	-2.677	
0.550	-3.142	-2.969	-2.799	-2.576	
0.555	-2.997	-2.843	-2.689	-2.485	
0.560	-2.870	-2.731	-2.591	-2.404	
0.565	-2.759	-2.632	-2.503	-2.329	
0.570	-2.659	-2.542	-2.423	-2.262	
0.575	-2.569	-2.462	-2.351	-2.199	
0.580	-2.488	-2.388	-2.284	-2.142	
0.585	-2.415	-2.321	-2.223	-2.089	
0.590	-2.347	-2.259	-2.167	-2.040	
0.595	-2.285	-2.202	-2.115	-1.993	
0.600	-2.227	-2.148	-2.066	-1.950	
0.610	-2.124	-2.053	-1.978	-1.872	
0.620	-2.034	-1.969	-1.900	-1.802	
0.630	-1.954	-1.894	-1.830	-1.739	
0.640	-1.883	-1.828	-1.768	-1.683	
0.650	-1.820	-1.768	-1.712	-1.632	
0.660	-1.762	-1.713	-1.660	-1.585	
0.670	-1.710	-1.663	-1.613	-1.541	
0.680	-1.662	-1.618	-1.570	-1.501	
0.690	-1.617	-1.576	-1.530	-1.464	
0.700	-1.577	-1.537	-1.493	-1.430	
0.710	-1.539	-1.500	-1.459	-1.398	
0.720	-1.503	-1.466	-1.426	-1.368	
0.730	-1.470	-1.435	-1.396	-1.340	
0.740	-1.440	-1.405	-1.368	-1.313	
0.750	-1.410	-1.377	-1.341	-1.288	
0.760	-1.383	-1.351	-1.316	-1.264	
0.770	-1.357	-1.326	-1.292	-1.242	
0.780	-1.333	-1.303	-1.269	-1.220	
0.790	-1.310	-1.280	-1.248	-1.200	
0.800	-1.288	-1.259	-1.227	-1.181	
0.850	-1.192	-1.167	-1.138	-1.096	
0.900	-1.115	-1.092	-1.066	-1.027	
0.950	-1.052	-1.030	-1.006	-0.970	
1.000	-0.998	-0.977	-0.955	-0.921	
1.050	-0.951	-0.932	-0.910	-0.878	
1.100	-0.911	-0.892	-0.872	-0.841	
1.150	-0.875	-0.857	-0.838	-0.808	
1.200	-0.843	-0.826	-0.807	-0.779	
1.250	-0.815	-0.798	-0.780	-0.752	
1.300	-0.789	-0.773	-0.755	-0.728	
1.350	-0.765	-0.750	-0.732	-0.706	
1.400	-0.744	-0.728	-0.712	-0.686	
1.450	-0.724	-0.709	-0.692	-0.668	
1.500	-0.705	-0.691	-0.675	-0.650	

# ALMOST UNBIASED RATIO ESTIMATES BASED ON INTERPENETRATING SUB-SAMPLE ESTIMATES

By M. N. MURTHY

and

N. S. NANJAMMA

*Indian Statistical Institute, Calcutta*

**SUMMARY.** In this paper a technique is developed to estimate the bias of an ordinary ratio estimate to a given degree of approximation on the basis of the interpenetrating sub-sample estimates. This estimate of the bias is used to correct the ratio estimate for its bias, thereby obtaining an almost unbiased ratio estimate.

## 1. INTRODUCTION

In large scale sample surveys, the method of ratio estimation is used to estimate various ratios. It is also used to estimate totals where supplementary information is available, since under certain circumstances usually met with, it is more efficient than the conventional methods of obtaining unbiased estimates. But a satisfactory treatment of the bias and error of a ratio estimate is not yet available. However different sampling and estimation procedures have been given which provide unbiased ratio estimates. In this paper, two different types of ratio estimates based on estimates obtained from  $n$  independent, and interpenetrating sub-samples have been compared from the points of view of bias (to the second degree of approximation) and mean square error (to the fourth degree of approximation). This study helps in obtaining an estimate of the bias of the ratio estimate, for any probability sampling design. Once the bias is estimated, the ratio estimate can be corrected to give an unbiased ratio estimate (unbiased to the second degree of approximation). The gain in precision of this unbiased ratio estimate as compared with the biased one has been studied.

The above results can be generalised to estimate the bias of a ratio estimate to any degree of approximation, using a series of ratio estimates based on a number of independent and interpenetrating sub-samples. These generalised results, for the particular case where the estimates of the variates in question are distributed in the bivariate normal form are given in sections 7 and 8 of this paper.

## 2. APPROXIMATIONS FOR THE BIAS AND MEAN SQUARE ERROR OF A RATIO ESTIMATE

Let  $\hat{y}$  and  $\hat{x}$  be unbiased estimates of  $y$  and  $x$ , the population totals of two characteristics, based on any probability sample.  $\hat{y}/\hat{x}$  can be considered as an estimate of the ratio  $R = y/x$ . This estimate is consistent but biased. Assuming that  $\frac{|\hat{x}-x|}{x} < 1$ , and neglecting terms of degree greater than two in the expansion of  $\left(1 + \frac{\hat{y}-y}{y}\right) \left(1 + \frac{\hat{x}-x}{x}\right)^{-1}$ , it can be shown that the bias of  $\hat{y}/\hat{x}$

$$B(\hat{y}/\hat{x}) = \frac{1}{x^2} [R \text{ var}(\hat{x}) - \text{cov}(\hat{x}, \hat{y})]. \quad \dots (2.1)$$



The mean square error of  $\hat{y}/\hat{x}$ , to the fourth degree of approximation, is

$$M(\hat{y}/\hat{x}) = R^2 \left\{ \left( \frac{\mu_{02}}{y^2} - \frac{2\mu_{11}}{xy} + \frac{\mu_{20}}{x^2} \right) + 2 \left( \frac{2\mu_{21}}{x^2y} - \frac{\mu_{12}}{xy^2} - \frac{\mu_{30}}{x^3} \right) + 3 \left( \frac{\mu_{22}}{x^2y^2} - \frac{2\mu_{31}}{x^3y} + \frac{\mu_{40}}{x^4} \right) \right\} \dots (2.2)$$

where

$$\mu_{ij} = \mathcal{E} [(\hat{x}-x)^i (\hat{y}-y)^j].$$

If the sample size is fairly large, the assumption  $\frac{|\hat{x}-x|}{x} < 1$  can be considered to be valid. Further  $x$  usually denotes the number of persons, or households, or some such characteristic for which we expect reliable estimates with a good design. For simple random sampling, a large number of empirical studies have shown that generally if the sample size is greater than 30, the assumption that  $\frac{|\hat{x}-x|}{x} < 1$  is valid; and that the contribution of the higher degree terms to the bias and variance of the ratio estimate will be negligible.

### 3. COMPARISON OF TWO DIFFERENT RATIO ESTIMATES

Let  $(y_i, x_i)$  be unbiased estimates of the population totals  $y$  and  $x$ , from the  $i$ -th independent interpenetrating sub-sample ( $i = 1, 2, \dots, n$ ). The following two ratio estimates can be considered as estimates of  $R = y/x$

$$(i) \quad R_1 = \frac{y_1 + y_2 + \dots + y_n}{x_1 + x_2 + \dots + x_n}$$

$$(ii) \quad R_n = \frac{1}{n} \left( \frac{y_1}{x_1} + \frac{y_2}{x_2} + \dots + \frac{y_n}{x_n} \right)$$

Applying result (2.1) to

$$R_1 = \left( \frac{\sum_{i=1}^n y_i/n}{\sum_{i=1}^n x_i/n} \right), \text{ we get the bias of } R_1,$$

$$B(R_1) = B_1 = \frac{1}{x^2} \left[ R \text{ var} \left( \frac{\sum x_i}{n} \right) - \text{cov} \left( \frac{\sum y_i}{n}, \frac{\sum x_i}{n} \right) \right]$$

since  $\mu_{11}(x_i, y_j) = 0$  for  $i \neq j$

$$\begin{aligned} &= \frac{1}{n^2 x^2} \sum_{i=1}^n \left\{ R \mu_{20}(x_i) - \mu_{11}(x_i, y_i) \right\} \\ &= \frac{1}{n^2} \left\{ \sum_{i=1}^n B \left( \frac{y_i}{x_i} \right) \right\}. \end{aligned} \dots (3.1)$$

Bias of  $R_n$ ,

$$B(R_n) = B_n = \frac{1}{n} \left\{ \sum_{i=1}^n B \left( \frac{y_i}{x_i} \right) \right\}. \dots (3.2)$$

Comparing (3.1) and (3.2), we note that the bias of  $R_n$  is  $n$  times that of  $R_1$ , to the second degree of approximation.

We now compare the mean square errors of  $R_1$  and  $R_n$  to the fourth degree of approximation, assuming that the sub-sample sizes are the same (as is the case generally) so that

$$B\left(\frac{y_i}{x_i}\right) = B$$

$$\mu_{rs}(x_i, y_i) = \mu_{rs}$$

and

$$M\left(\frac{y_i}{x_i}\right) = M \text{ for all } i.$$

By applying result (2.2) to  $R_1$  and simplifying, we obtain,

$$\begin{aligned} M(R_1) = M_1 &= \frac{R^2}{n} \left\{ \left( \frac{\mu_{02}}{y^2} + \frac{\mu_{20}}{x^2} - \frac{2\mu_{11}}{xy} \right) + \frac{2}{n} \left( \frac{2\mu_{21}}{x^2y} - \frac{\mu_{30}}{x^3} - \frac{\mu_{12}}{xy^2} \right) + \right. \\ &\quad \left. + \frac{3}{n^2} \left( \frac{\mu_{40}}{x^4} + \frac{\mu_{22}}{x^2y^2} - \frac{2\mu_{31}}{x^3y} \right) + \frac{3(n-1)}{n^2} \left( \frac{3\mu_{20}^2}{x^4} + \frac{\mu_{20}\mu_{02}}{x^2y^2} + \frac{2\mu_{11}^2}{x^2y^2} - \frac{6\mu_{20}\mu_{11}}{x^3y} \right) \right\} \\ &= \frac{M}{n} - \frac{n-1}{n^2} A \end{aligned} \quad \dots (3.3)$$

where

$$\begin{aligned} A = R^2 \left[ 2 \left( \frac{2\mu_{21}}{x^2y} - \frac{\mu_{30}}{x^3} - \frac{\mu_{12}}{xy^2} \right) + \frac{3(n+1)}{n} \left( \frac{\mu_{40}}{x^4} + \frac{\mu_{22}}{x^2y^2} - \frac{2\mu_{31}}{x^3y} \right) - \right. \\ \left. - \frac{3}{n} \left( \frac{3\mu_{20}^2}{x^4} + \frac{\mu_{20}\mu_{02}}{x^2y^2} + \frac{2\mu_{11}^2}{x^2y^2} - \frac{6\mu_{20}\mu_{11}}{x^3y} \right) \right], \end{aligned}$$

and

$$M(R_n) = M_n = \mathcal{E} (R_n - R)^2 = \mathcal{E} \left[ \frac{1}{n} \sum_{i=1}^n \left( \frac{y_i}{x_i} - R \right) \right]^2$$

$$= \frac{1}{n^2} \sum_{i=1}^n \mathcal{E} \left( \frac{y_i}{x_i} - R \right)^2 + \frac{1}{n^2} \sum_{i \neq j}^n B \left( \frac{y_i}{x_i} \right) B \left( \frac{y_j}{x_j} \right)$$

$$= \frac{1}{n^2} \sum_{i=1}^n M \left( \frac{y_i}{x_i} \right) + \frac{1}{n^2} \sum_{i \neq j}^n B \left( \frac{y_i}{x_i} \right) B \left( \frac{y_j}{x_j} \right) = \frac{M}{n} + \frac{(n-1)B^2}{n} \quad \dots (3.4)$$

from (3.3) and (3.4)  $M_n = M_1 + \frac{n-1}{n^2} A + \frac{n-1}{n} B^2$ .



Comparison of  $M_n$  and  $M_1$  is difficult in general. Hence it is assumed that  $\hat{x}$  and  $\hat{y}$  are distributed in the bivariate normal form in which case the bias and mean square error of  $\hat{y}/\hat{x}$  reduce to

$$B = Rc_x(c_x - \rho c_y)$$

$$M = R^2[(c_y^2 - 2\rho c_x c_y + c_x^2)(1 + 3c_x^2) + 6c_x^2(c_x - \rho c_y)^2]$$

$$\text{Further } A = 3R^2c_x^2[(c_y^2 - 2\rho c_x c_y + c_x^2) + 2(c_x - \rho c_y)^2], \text{ which is } > 0$$

$$\text{where } c_x^2 = \frac{\mu_{20}}{x^2}, c_y^2 = \frac{\mu_{02}}{y^2},$$

and  $\rho$  = correlation coefficient between  $\hat{x}$  and  $\hat{y}$ .

$\therefore$  The mean square error of  $R_n$  is greater than that of  $R_1$ . Thus  $R_1$  is better than  $R_n$  from the considerations of both bias and mean square error.

#### 4. ESTIMATION OF THE BIAS OF THE RATIO ESTIMATE

An unbiased estimate of the bias of the ratio estimate to the second degree of approximation, is given below.

$$\mathcal{E}(R_1) = R + B_1$$

$$\mathcal{E}(R_n) = R + B_n$$

$$\therefore \mathcal{E}(R_n - R_1) = B_n - B_1.$$

But

$$B_n = nB_1$$

$$\therefore \mathcal{E}(R_n - R_1) = (n-1)B_1.$$

$\therefore \hat{B}_1 = \frac{R_n - R_1}{n-1}$  is an unbiased estimate of  $B_1$ , the bias of  $R_1$  to the second degree of approximation.

The variance of the estimate of bias of  $R_1$  is given by

$$V(\hat{B}_1) = \frac{1}{(n-1)^2} (V_1 + V_n - 2\rho_{R_1 R_n} \sqrt{V_1 V_n})$$

where

$$V_1 = \text{Variance of } R_1 = M_1 - B_1^2$$

$$V_n = \text{Variance of } R_n = M_n - B_n^2$$

and

$$\rho_{R_1 R_n} = \text{Correlation coefficient of } R_1 \text{ and } R_n.$$

$\therefore$  using (3.1), (3.2), (3.3) and (3.4), we get

$$V_1 = V_n + \frac{n-1}{n^2} (B^2 - A)$$

$$\therefore V(\hat{B}_1) = \frac{V_n}{(n-1)^2} (\alpha^2 - 2\rho_{R_1 R_n} \alpha + 1) \quad \dots (4.1)$$

where

$$\alpha^2 = 1 + \frac{n-1}{n^2} \frac{B^2 - A}{V_n}.$$

# RATIO ESTIMATES BASED ON INTERPENETRATING SUB-SAMPLE ESTIMATES

If  $\hat{x}$  and  $\hat{y}$  are bivariate normally distributed

$$A - B^2 = 3c_x^2 R^2 [(c_y^2 - 2\rho c_x c_y + c^2 x) + \frac{5}{3} (c_x - \rho c_y)^2]$$

which is greater than or equal to 0.

It follows that  $\alpha^2 \leq 1$  and therefore  $V_1 \leq V_n$ . Thus it may be observed that  $R_1$  is a better estimate than  $R_n$  from the point of view of bias, mean square error as well as variance.

The expression for the correlation coefficient between  $R_1$  and  $R_n$  to the fourth degree of approximation is

$$\rho_{R_1 R_n} = \frac{n \left\{ \frac{n^2 + 3n + (1 + \rho)(n + 1)c^2}{(n + c^2)^2} - \frac{n + 1}{n} - \frac{c^2(1 - \rho)}{n} \right\}}{(11c^2 - 5c^2\rho + 2)^{\frac{1}{2}}(11c^2 - 5c^2\rho + 2n)^{\frac{1}{2}}} \quad \dots \quad (4.2)$$

under the assumption that

(a)  $\hat{x}$  and  $\hat{y}$  are bivariate normally distributed with the same coefficient of variation ( $c_x = c_y = c$ ) and

$$(b)^1 \quad (x_1^2 + x_1 x_2 + \dots + x_1 x_n) < 2[V(\hat{x}) + n x^2].$$

In the above expression  $\rho$  stands for the correlation coefficient of  $\hat{x}$  and  $\hat{y}$ . If  $c^2$  is small,  $\rho_{R_1 R_n}$  will be nearly equal to 1.

The coefficient of variation of the estimate of bias may be large; still it may be possible to get a ratio estimate corrected for its bias which is more efficient than the biased one.

## 5. AN ALMOST UNBIASED RATIO ESTIMATE

Since we have obtained an estimate of bias of  $R_1$ , that can be used to correct  $R_1$  for its bias, and we get an almost unbiased ratio estimate.

$$R_c = \left( \frac{n R_1 - R_n}{n - 1} \right)$$

---

<sup>1</sup> This assumption was necessary to derive the  $\mathcal{E}(R_1 R_n)$ , for

$$\rho_{R_1 R_n} = \frac{\mathcal{E}(R_1 R_n) - \mathcal{E}(R_1)\mathcal{E}(R_n)}{(V(R_1)V(R_n))^{\frac{1}{2}}}$$

$$\mathcal{E}(R_1 R_n) = \mathcal{E} \left( \frac{y_1^2 + y_1 y_2 + \dots + y_1 y_n}{x_1^2 + x_1 x_2 + \dots + x_1 x_n} \right)$$

$$= \frac{V(\hat{y}) + n y^2}{V(\hat{x}) + n x^2} + \frac{(V(\hat{y}) + n y^2)V(\varepsilon') - (V(\hat{x}) + n x^2) \text{cov.}(\varepsilon, \varepsilon')}{(V(\hat{x}) + n x^2)}$$

where

$$\varepsilon = (y_1^2 + y_1 y_2 + \dots + y_1 y_n) - (V(\hat{y}) + n y^2)$$

$$\varepsilon' = (x_1^2 + x_1 x_2 + \dots + x_1 x_n) - (V(\hat{x}) + n x^2).$$



We say it is an 'almost' unbiased estimate because it is unbiased only to the second degree of approximation. The variance of the corrected estimate is

$$\begin{aligned} V(R_c) &= \frac{1}{(n-1)^2} \left( n^2 V_1 + V_n - 2n\rho_{R_1 R_n} \sqrt{V_1 V_n} \right) \\ &= \frac{V_n}{(n-1)^2} \left( n^2 \alpha^2 - 2n\rho_{R_1 R_n} \alpha + 1 \right) \end{aligned} \quad \dots (5.1)$$

The gain in precision due to using  $R_{1c}$  instead of  $R_1$  is given by

$$G(R_c) = \frac{M_1 - V(R_c)}{M_1} = 1 - \frac{n^2 \alpha^2 - 2n\rho_{R_1 R_n} \alpha + 1}{(n-1)^2 (\alpha^2 + z^2)} \quad \dots (5.2)$$

where  $z^2 = \frac{B^2}{n^2 V_n} = \frac{B^2}{nV}$ ,  $B$  and  $V$  being the bias and variance of the ratio estimate based on one sub-sample. It may be noted that

$$\frac{|B|}{V^{\frac{1}{2}}} < c_x.$$

where  $c_x$  is the coefficient of variation of the estimate  $\hat{x}$  based on one sub-sample. If the sub-sample size is large  $c_x$  will be small. Hence  $z^2$  can be neglected. It is to be noted that neglecting  $z^2$  does not amount to neglecting bias. The gain in precision can be written as

$$G(R_c) = 1 - \frac{n^2 \alpha^2 - 2n\rho_{R_1 R_n} \alpha + 1}{(n-1)^2 \alpha^2} \quad \dots (5.3)$$

Further the expression in (5.2) is greater than that in (5.3).

$$G(R_c) > 0, \text{ if } (n-1)^2 \alpha^2 - (n^2 \alpha^2 - 2n\rho_{R_1 R_n} \alpha + 1) > 0$$

$$\text{i.e., if } (2n-1)\alpha^2 - 2n\rho_{R_1 R_n} \alpha + 1 < 0.$$

which will be true if  $\alpha$  lies between the roots of the equation

$$(2n-1)\alpha^2 - 2n\rho_{R_1 R_n} \alpha + 1 = 0 \quad \dots (5.4)$$

$$\text{(i.e.) if } \alpha \text{ lies between } \frac{n\rho_{R_1 R_n} \pm (n^2 \rho_{R_1 R_n}^2 - 2n + 1)^{\frac{1}{2}}}{(2n-1)}.$$

For given values of  $\alpha$  and  $\rho_{R_1 R_n}$ , the minimum value of  $n$  which makes  $G(R_c) > 0$  is given by

$$n = \left[ \frac{(1-\alpha^2)}{2\alpha(\rho_{R_1 R_n} - \alpha)} \right] + 1 \quad \dots (5.5)$$

# RATIO ESTIMATES BASED ON INTERPENETRATING SUB-SAMPLE ESTIMATES

It can be seen that  $G(R_c)$  will be positive only if  $\rho_{R_1 R_n} > \alpha$ . Further for given values of  $\alpha$  and  $\rho_{R_1 R_n}$  where  $\rho_{R_1 R_n} > \alpha$ , the value of  $n$  which maximises the gain is

$$n = \frac{(1 - \rho_{R_1 R_n} \alpha)}{\alpha (\rho_{R_1 R_n} - \alpha)} \quad \dots (5.6)$$

For given values of  $n$  and  $\rho_{R_1 R_n}$  the value of  $\alpha$  which maximises the gain is

$$\alpha = \frac{1}{n \rho_{R_1 R_n}}.$$

A table showing for given values of  $\rho_{R_1 R_n}$  and  $\alpha$ , the minimum value of  $n$  required to make the gain positive, the optimum  $n$  and the maximum gain are given below.

MINIMUM AND MAXIMUM VALUES OF  $G(R_c)$  WITH THE CORRESPONDING VALUES OF  $n$  FOR DIFFERENT  $\rho_{R_1 R_n}$  AND  $\alpha$  WHERE  $\rho_{R_1 R_n} > \alpha$ .

sl. no.	$\alpha$	$\rho_{R_1 R_n}$	minimum		maximum	
			$n$	$G(R_c)$	$n$	$G(R_c)$
(0)	(1)	(2)	(3)	(4)	(5)	(6)
1	0.6	0.7	6	0.0089	10	0.0192
2		0.8	3	0.0556	4	0.0988
3		0.9	2	0.0889	3	0.3056
4	0.7	0.8	4	0.0113	7	0.0266
5		0.9	2	0.1020	3	0.1684
6	0.8	0.9	3	0.0469	4	0.0486

## 6. EMPIRICAL STUDY

In section 5 we have discussed the efficiency of the corrected estimate  $R_c$  as compared to that of  $R_1$ . There it has been pointed out that under certain circumstances  $R_c$  will be a better estimate of  $R$  than  $R_1$ . In this section we give an example where the variance of  $R_c$  turned out to be less than the mean square error of  $R_1$ .

The data for this study consist of the village-wise figures for the number of households and the number of persons attending the village market for a sample of 300 villages scattered over a wide region. Treating this as the population two samples of size 30 villages are drawn systematically with independent random starts to estimate the number of persons going to market per household. From these two sub-samples the estimates  $R_1$  and  $R_n$  are calculated. Then the corrected estimate is found by estimating the bias of  $R_1$ .

For the purpose of this study, all the possible pairs of systematic samples are enumerated. Then for each of the pairs  $R_1$ ,  $R_n$ , and  $R_n - R_1$  are calculated. The variance of  $R_c$ ,



the corrected estimate and the mean square error of  $R_1$  are determined. The results are given below.

$$\text{Population ratio} = 7.6857$$

$$E(R_1) = 7.9211 \text{ and } E(R_n) = 8.1401$$

$$B(R_1) = 0.2354 \text{ and } B(R_n) = 0.4544$$

It is to be noted that  $B_1$  is almost half of  $B_n$ . This may be taken as indicating that second degree approximation is good enough. The variance of  $R_c$  and mean square error of  $R_1$  are

$$V(R_c) = 8.9992 \text{ and } M(R_1) = 9.6144$$

$$\rho_{R_1 R_n} = 0.9856 \text{ and } \alpha = 0.8871$$

$$G(R_c) = 6.4\%$$

#### 7. COMPARISON OF A SERIES OF RATIO ESTIMATES

When  $n$ , the number of independent and interpenetrating sub-samples is a multiple of 2, 3, ... and  $k$ , we can construct the following series of ratio estimates.

$$R_m = \frac{1}{m} \left\{ \frac{\sum_{i=1}^{\frac{n}{m}} y_i}{\sum_{i=1}^{\frac{n}{m}} x_i} + \frac{\sum_{i=\frac{n}{m}+1}^{2\frac{n}{m}} y_i}{\sum_{i=\frac{n}{m}+1}^{2\frac{n}{m}} x_i} + \dots + \frac{\sum_{i=(m-1)\frac{n}{m}+1}^n y_i}{\sum_{i=(m-1)\frac{n}{m}+1}^n x_i} \right\}$$

where  $m = 1, 2, 3, \dots, k, n$ .

$$= \frac{1}{m} \left\{ \sum_{j=1}^m (R_{n/m})_j \right\} \text{ where } (R_{n/m})_j = \frac{\sum_{i=(j-1)\frac{n}{m}+1}^{j\frac{n}{m}} y_i}{\sum_{i=(j-1)\frac{n}{m}+1}^{j\frac{n}{m}} x_i}$$

It may be noted that there are  $\frac{n!}{m! \left\{ \left( \frac{n}{m} \right)! \right\}^m}$  different ways of partitioning the  $n$  sub-samples into  $m$  partitions each containing  $n/m$  sub-samples.

In practice the situation may arise where  $(\hat{x}, \hat{y})$  are approximately distributed in the bivariate normal form. In such a case we may make use of the properties of the bivariate normal distribution for writing down the expressions for the bias and mean square error of

the ratio estimate. It is to be noted that the infinite series for the bias and mean square error are divergent. As has been rightly pointed out by Cramér and Kendall in their books, in statistical practice one is interested not so much in the convergence properties of the infinite series representing a function but in finding out whether the first few terms of that series will give a good approximation to the function.

Naturally in a finite population where the estimate  $\hat{x}$  does not take the value 0, the bias and the mean square error of the ratio estimate  $\hat{y}/\hat{x}$  will be finite quantities. The formal expressions for the bias and mean square error in terms of infinite series under the assumption of bivariate normality are considered here. The problem as to how many terms are to be taken to obtain a desired degree of approximation in different situations is yet to be fully investigated. So in discussing the bias, only terms of degree greater than  $2k$  are neglected where  $k$  is any finite number.

$$\text{Bias of } \frac{\hat{y}}{\hat{x}} \text{ is given by } B(\hat{y}/\hat{x}) = R(c_x - \rho c_y) \sum_{j=1}^k \frac{(2j)!}{2^j j!} c_x^{2j-1} = \sum_{j=1}^k A_j \quad \dots (7.1)$$

$$\text{where } A_j = R(c_x - \rho c_y) \frac{(2j)!}{2^j j!} c_x^{2j-1}.$$

Mean square error of  $\frac{\hat{y}}{\hat{x}}$  to the fourth degree of approximation is given by

$$M\left(\frac{\hat{y}}{\hat{x}}\right) = R^2\{(1 + 3c_x^2)(c_y^2 - 2\rho c_x c_y + c_x^2) + 6c_x^2(c_x - \rho c_y)^2\}. \quad \dots (7.2)$$

$$\text{Bias of } R_m, B(R_m) = B_m = \left(\frac{m}{n}\right)^{\frac{1}{2}} R(c_x - \rho c_y) \sum_{j=1}^k \frac{(2j)!}{2^j j!} \frac{c_x^{2j-1}}{\left(\frac{n}{m}\right)^{j-\frac{1}{2}}} = \sum_{j=1}^k \frac{m^j}{n^j} A_j = \sum_{j=1}^k m^j \eta_j \quad \dots (7.3)$$

$$\text{where } \eta_j = \frac{A_j}{n^j}.$$

From (7.3) it follows that the biases of the series of ratio estimates have the same sign, and the absolute magnitude of the bias increases with  $m$ . From (7.2) the mean square error of  $R_m$  to the fourth degree of approximation is given by

$$\begin{aligned} M(R_m) = M_m &= \frac{R}{n} (c_y^2 - 2\rho c_x c_y + c_x^2) + \frac{m}{n^2} A + \frac{m(m-1)}{n^2} B^2 \\ &= \frac{M-A}{n} + \frac{m}{n^2} A + \frac{m(m-1)}{n^2} B^2 \end{aligned}$$

where  $A$ ,  $B$  and  $M$  have been defined in section 3. Since  $A > 0$ ,  $M_m$  is an increasing function of  $m$ .

Further  $M_m - M_{m-1} = \frac{A}{n^2} + \frac{2(m-1)}{n^2} B^2$  which also increases as  $m$  increases.



# 8. ESTIMATION OF THE BIAS OF A SERIES OF RATIO ESTIMATES

The bias of  $R_m$  to the  $(2k)$ -th degree of approximation can be estimated as given below, from  $n$  independent and interpenetrating sub-samples, provided<sup>1</sup>  $n$  is a multiple of 2, 3, 4 ... and  $k$ , where  $\hat{x}$  and  $\hat{y}$  are distributed in the bivariate normal form. (The bias, when  $\hat{x}$  and  $\hat{y}$  are not bivariate normally distributed can also be estimated by adopting a similar procedure).

The bias of  $R_m$  to the  $2k$ -th degree of approximation is given by

$$B = \left( \sum_{i=1}^k m^j \eta_j \right) \quad m = 1, 2, \dots, k, n \quad [\text{see (7.3)}] \quad \dots \quad (7.4)$$

$$\mathcal{E}(R_m) = R + B_m$$

$$\therefore \mathcal{E}(R_m - R_1) = B_m - B_1 = \sum_{j=1}^k (m^j - 1) \eta_j \quad \dots \quad (7.5)$$

Let

$$D_m = R_m - R_1$$

From (7.5)

$$\mathcal{E}(D) = \begin{pmatrix} \eta \\ 1 \times k \end{pmatrix} \quad \begin{pmatrix} \wedge \\ 1 \times k \end{pmatrix}$$

where

$$D = (D_2, D_3, \dots, D_k, D_n)$$

$$\eta = (\eta_1, \eta_2, \dots, \eta_k)$$

$$\Lambda = \begin{pmatrix} 2-1 & 3-1 & \dots & k-1 & n-1 \\ 2^2-1 & 3^2-1 & \dots & k^2-1 & n^2-1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 2^k-1 & 3^k-1 & \dots & k^k-1 & n^k-1 \end{pmatrix}$$

$$\therefore \eta = \mathcal{E}(D) \wedge^{-1}$$

From (5.6)

$$\begin{pmatrix} B \\ 1 \times k \end{pmatrix} = \begin{pmatrix} \eta \\ 1 \times k \end{pmatrix} \begin{pmatrix} \wedge + \epsilon \\ k \times k \end{pmatrix}$$

where

$$B = (B_2, B_3, \dots, B_k, B_n)$$

and  $\epsilon$  is a  $(k \times k)$  matrix whose elements are all equal to 1.

$$\therefore B = \mathcal{E}[D] + \mathcal{E}[D] \wedge^{-1} \epsilon$$

But

$$\wedge^{-1} \epsilon = \begin{pmatrix} s_1 & s_1 & \dots & s_1 \\ s_2 & s_2 & \dots & s_2 \\ \vdots & \vdots & \ddots & \vdots \\ s_k & s_k & \dots & s_k \end{pmatrix}$$

where  $s_m$  is the sum of the elements in the  $m$ -th row of  $\wedge^{-1}$ .

$\therefore$  An estimate of  $(B)$ ,  $(\hat{B}) = D + D \wedge^{-1} \epsilon = D + D[S](1)$

where

$$[S] = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_k \end{pmatrix} \quad \text{and } (1) = (1, 1, \dots, 1)$$

$$\therefore \hat{B}_m = \sum_j D_j S_{j-1} + D_m \quad \text{where } j = 2, 3, \dots, k, n, \text{ and } s_{n-1} = s_k \quad \dots \quad (7.6)$$

Hence the corrected estimate is given by  $R_c = R_m - \hat{B}_m$ .

<sup>1</sup> This condition is only sufficient but not necessary. The estimate of bias and hence the corrected estimate may be obtained even if it is not a multiple of 2, 3, ...  $k$  by considering the series of ratio estimates defined over over-lapping partitions of the  $n$  sub-samples.

# RATIO ESTIMATES BASED ON INTERPENETRATING SUB-SAMPLE ESTIMATES

Particular cases :

$$(1) \quad k = 1 : R_1 = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}, R_n = \frac{1}{n} \sum_{i=1}^n \left( \frac{y_i}{x_i} \right)$$

$$\therefore B_m = m\eta_1 = \frac{m}{n} A_1 \text{ where } m = 1, n.$$

$$\text{Since } \Lambda = (n-1), \Lambda^{-1} = \frac{1}{n-1} \text{ and } [S] = \frac{1}{n-1}$$

$$\therefore B_n = (R_n - R_1) + (R_n - R_1) \frac{1}{n-1} = \frac{n}{n-1} (R_n - R_1)$$

$$\text{But } B_n - B_1 = \mathcal{E} (R_n - R_1)$$

$$\therefore \hat{B}_1 = \hat{B}_n - (R_n - R_1) = \frac{n}{n-1} (R_n - R_1) - (R_n - R_1) = \frac{R_n - R_1}{n-1}$$

$$(2) \quad k = 2 : R_1 = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}, R_2 = \left\{ \frac{\sum_{i=1}^{n/2} y_i}{\sum_{i=1}^{n/2} x_i} + \frac{\sum_{i=(n/2)+1}^n y_i}{\sum_{i=(n/2)+1}^n x_i} \right\} \text{ and } R_n = \frac{1}{n} \sum_{i=1}^n \left( \frac{y_i}{x_i} \right)$$

where  $n$  is a multiple of 2.

$$\text{Here } B_m = m\eta_1 + m^2\eta_2 = \frac{m}{n} A_1 + \frac{m^2}{n^2} A_2; m = 1, 2 \text{ and } n.$$

$$\Lambda = \begin{pmatrix} 2-1 & n-1 \\ 2^2-1 & n^2-1 \end{pmatrix} = \begin{pmatrix} 1 & n-1 \\ 3 & n^2-1 \end{pmatrix}$$

$$\therefore \Lambda^{-1} = \frac{1}{(n-1)(n-2)} \begin{pmatrix} n^2-1 & -(n-1) \\ -3 & 1 \end{pmatrix}$$

$$\therefore s_1 = \frac{n}{n-2}, s_2 = \frac{-2}{(n-1)(n-2)}$$

$$\therefore (\hat{B}_2, \hat{B}_n) = (R_2 - R_1, R_n - R_1) + (R_2 - R_1, R_n - R_1) \begin{pmatrix} \frac{n}{n-2} \\ -2 \\ \frac{-2}{(n-1)(n-2)} \end{pmatrix} (1, 1)$$



$$\therefore \hat{B}_2 = (R_2 - R_1) + \frac{n}{n-2} (R_2 - R_1) - \frac{2(R_n - R_1)}{(n-1)(n-2)}$$

$$= \frac{-2n}{n-1} R_1 + \frac{2(n-1)}{n-2} R_2 - \frac{2R_n}{(n-1)(n-2)}$$

similarly 
$$\hat{B}_n = \frac{-2n}{n-1} R_1 + \frac{n}{n-2} R_2 + \frac{n(n-3)R_n}{(n-1)(n-2)}$$

and 
$$\hat{B}_1 = -\frac{(n+1)}{n-1} R_1 + \frac{nR_2}{n-2} - \frac{2R_n}{(n-1)(n-2)}$$

Hence the almost unbiased estimate in this case is

$$R_c = \frac{2nR_1}{n-1} - \frac{nR_2}{n-2} + \frac{2R_n}{(n-1)(n-2)}$$

It may be noted that the results given above will also be obtained when an estimate of the bias to the third degree of approximation is considered, in the case when  $\hat{x}$  and  $\hat{y}$  are not distributed in the bivariate normal form.

The authors wish to thank Prof. D. B. Lahiri for his guidance in preparing this paper.

#### REFERENCES

- CRAMÉR, H. (1946): *Mathematical Methods of Statistics*. Princeton University Press.
- FIELLER, E. C. (1932): The distribution of the index in a normal bivariate population. *Biometrika*, 24, 428-440.
- KENDALL, M. G. (1948): *Advanced Theory of Statistics*, Vol. I., Charles Griffin & Co.
- SUKHATME, P. V. (1953): *Sampling Theory of Surveys with Applications*. The Iowa State College Press.

*Paper received : November, 1956.*

*Revised : March, 1959.*

# PRECISION IN THE CONSTRUCTION OF COST OF LIVING INDEX NUMBERS

By K. S. BANERJEE

*State Statistical Bureau, West Bengal*

**SUMMARY.** In the construction of cost of living index (consumers' price index) numbers, consumption items are usually grouped into composite commodities. A criterion has been developed showing how best the grouping of items could be done. That non-judicious grouping might lead to serious errors has also been demonstrated.

## 1. INTRODUCTION

Although much attention has been drawn to the problem of securing the True Index (TI) in the context of constructing Cost of Living Index (CLI) numbers, and formulae evolved [Banerjee (1956c), Frisch (1936), Konüs (1939) and Wald (1939)] for the purpose of constructing the True Index, these formulae do not appear to have been used much in practice. In actual practice, however, Laspeyres' base-weighted formula continues to be widely used for approximating the CLI, although it is known to over-estimate the index.

## 2. PRECISION NEGLECTED

Whereas it was necessary to construct the True Index in the precise estimation of CLI, and whereas, instead, Laspeyres' formula is being used at the cost of precision, it would, at least, only be reasonable to make sure that Laspeyres' Index be precisely calculated. This aspect of precision does not appear to have been paid the attention it deserves, so much so that it sometimes causes an embarrassment, when different organisations, while calculating the CLI for the same area and the same economic stratum of population, come out with different figures for the same index. Difference in the figures for the same index could have been appreciated, if the coverage (the sample, or the way the sample is selected) and the error of estimation were made available. In absence of such information, controversies arise causing difficulty at administrative levels. With a view to systematising the study, the concept of standard error in index number calculation was introduced in an earlier note (Banerjee, 1956a) where it was shown that it would be possible to calculate the standard error for an estimated CLI under certain assumptions.<sup>1</sup>

## 3. PURPOSE OF THIS NOTE

The purpose of this note is to show the extent of error which might creep in through a non-judicious computation of Laspeyres' Index and to suggest measures of precaution to be taken for precision. It is also to demonstrate incidentally some principles which would serve as a guide for calculating Laspeyres' Index on minimum price collection and minimum computation.

---

<sup>1</sup> These assumptions have later been generalised leading to the same form of error variance as hereinafter indicated.



The principles which have been demonstrated here will have their application in general in any index number formula, which, or a part of which, is reducible to the form of a weighted average of relatives.

#### 4. LASPEYRES' FORMULA

Laspeyres' formula,  $100 \frac{\sum_{i=1}^N p_{1i} q_{0i}}{\sum_{i=1}^N p_{0i} q_{0i}}$ , is usually adopted in routine practice in the reduced form,  $\sum_i r_i w_i$ , where  $r_i (= 100 p_{1i}/p_{0i})$  is the price relative of the  $i$ -th consumption item expressed in percentage,  $w_i (= p_{0i} q_{0i} / \sum_{i=1}^N p_{0i} q_{0i})$  the weight of the  $i$ -th item which is known and is determined from family budgets, and  $\sum_i w_i = 1$ . The consumption items are usually grouped under major groups of consumption, and within each major group the items are again grouped into sub-groups which, in turn, may either be composite commodities or singular items. For each such sub-group, a price relative is obtained, and such price relatives are averaged with the corresponding weights.

The calculation of the index is generally completed by two stages. It is first calculated for a major group, and then the indexes for the major groups are combined into the overall index.

Usually, a sub-group is also a composite commodity consisting of numerous, though finite, constituent items. In that case, the calculation has to be extended to the third stage, beginning with the index for such a sub-group (composite commodity).

Without loss of generality, however, the calculation of the index may be considered to be a two-stage one; that is, the index will be completed first for a composite commodity (sub-group), and then the indexes for the composite commodities (sub-groups) will be combined into the overall index. In that case, Laspeyres' formula will take the form,

$$\sum_{i=1}^g \sum_{j=1}^{N_i} r_{ij} w_{ij}, \quad \begin{matrix} i = 1, 2, \dots, g \\ j = 1, 2, \dots, N_i \end{matrix} \quad \dots \quad (4.1)$$

where  $r_{ij}$  and  $w_{ij}$  are respectively the price relative and the weight of the  $j$ -th constituent item of the  $i$ -th composite commodity,  $\sum_i \sum_j w_{ij} = 1$ ,  $\sum_j w_{ij} = w_i$ ,  $\sum_i w_i = 1$  and  $\sum_i N_i = N$ .

#### 5. A PRACTICE IN VOGUE IN THE TREATMENT OF COMPOSITE COMMODITY

The weights of the individual constituent items of a composite commodity are not known in practice, as it becomes impracticable to determine the individual weights from a Family Budget Enquiry. What is afforded by a Family Budget Enquiry is the *total* weight for all the constituent items constituting the composite commodity.

## PRECISION IN THE CONSTRUCTION OF COST OF LIVING INDEX NUMBERS

$r_{ij}$  of (4.1) is, therefore, known, but not  $w_{ij}$ . This absence of knowledge of  $w_{ij}$  brings in the difficulty in the precise estimation of the index and, for the matter of that, is responsible for bringing into being divergent practices in the actual computation of the index.

To meet this difficulty of not having the knowledge of  $w_{ij}$ , the practice is to calculate some sort of a price relative for the composite commodity as a whole and then to have it weighted with the total weight for the entire composite commodity. The price relative used, under one such practice,<sup>1</sup> is the relative of the average prices in the base period, 0, and in the given period, 1, the average price of the composite commodity being *defined as the price averaged with the quantities sold of the constituent items in the market*.

With such a definition for the average price, the treatment of the composite commodity will be in agreement, subject to sampling fluctuations (Banerjee, 1956b), with the requirement of Laspeyres' Index. But there are certain assumptions involved in the practice which may not be always realisable for all types of composite commodities. The assumptions are :

(i) Acceptance of the definition for the average price of a composite commodity in the way it has been framed above.

(ii) Existence of the same supply pattern (relative supply) both in the base period (0) and the period of comparison (1).

Here relative supply of a constituent item has been taken to mean the proportion of its supply to the total supply of the composite commodity.

Assumption (i) in a given period could be accepted as reasonable, only if the relative supply of the constituent items would remain same during the period. While this might hold good in respect of many composite commodities, there may be some composite commodities where it might be violently wrong to make this assumption. If for those composite commodities, only the cheaper of the constituent items appear, for some reason or other, in the market in any period for sale, the average price of the composite commodity would come out as less than what actually it should be, and *vice versa*.

If validity of assumption (ii) could be accepted at all, it could perhaps be accepted to hold good during shorter intervals. At least, the probability of the assumption holding good during shorter intervals of time would be higher than that during wider intervals. That it would be so is a limitation of assumption (ii).

### 6. A REASONABLE PROCEDURE AND OUTLOOK

A reasonable procedure would be to calculate the price relatives, for some  $n_i$  constituent items out of a total of  $N_i$  constituting the composite commodity, and to take an arithmetic mean of the  $n_i$  price relatives,  $r_{ik}$  ( $k = 1, 2, \dots, n_i$ ) so as to calculate the index as

$\sum_{i=1}^g \bar{r}_i^* w_i$ , where  $\bar{r}_i^* = \frac{1}{n_i} \sum_{k=1}^{n_i} r_{ik}$ . The implications of this practice may be indicated as follows:

---

<sup>1</sup> Reference to some other practices has been made in an earlier note. [(Banerjee) : *Bull. Cal. Stat. Ass.*, 7 No. 25, 1956, pp. 35-40].



Let  $\rho_i$  be the correlation coefficient between  $r_{ij}$  and  $w_{ij}$ , and  $w'_{ij} = w_{ij}/w_i$ . Then, for  $i$ -th composite commodity, we have

$$\sum_j r_{ij} w'_{ij} = \sum_j r_{ij} / N_i + N_i \rho_i \sigma_{r_{ij}} \sigma_{w'_{ij}},$$

$$\text{or,} \quad \sum_j r_{ij} w_{ij} = \bar{r}_i w_i + N_i \rho_i \sigma_{r_{ij}} \sigma_{w_{ij}}, \quad \dots (6.1)$$

where  $\bar{r}_i = \sum_{j=1}^{N_i} r_{ij} / N_i$ . If  $\rho_i = 0$ , we shall have  $\sum_j r_{ij} w_{ij} = \bar{r}_i w_i$ . Under this condition, formula (4.1) would reduce to  $\sum_{i=1}^g \bar{r}_i w_i$ . The condition,  $\rho_i = 0$ , therefore, dispenses with the necessity for having to know the individual weights,  $w_{ij}$ .

The error variance of the estimate,  $\sum_i \bar{r}_i^* w_i$ , may be derived in the form,

$$\sum_i \frac{N_i - n_i}{N_i - 1} \frac{\sigma_i^2 w_i^2}{n_i}, \quad \dots (6.2)$$

where  $\sigma_i^2$  is the variance of  $r_{ij}$  within  $i$ .

It has to be remembered in this context that the individual weights of the constituent items can be pooled, only if the correlation coefficient between the price relatives and the weights of the constituent items is zero.

If  $\rho_i$  is not equal to zero, this practice will lead to an erroneous result. If  $\rho_i$  is either positive, or negative for all  $i$ 's, the errors will be additive bringing in a wide departure from what is being estimated. If, however, some of  $\rho_i$ 's are positive and some negative, the errors will partly cancel out and, as a result, the magnitude of the added errors will be less. If  $\rho_i$  is not equal to zero, its contribution to the error will be  $N_i \rho_i \sigma_{r_{ij}} \sigma_{w_{ij}} = N_i b_i \sigma_{r_{ij}}$ , where  $b_i$  is the regression coefficient of  $w_{ij}$  on  $r_{ij}$ . Therefore, if each of the  $\rho_i$ 's is not zero individually, we should have  $\sum_{i=1}^g N_i b_i (\sigma_i)^2 = 0$  so that no error is made.

It appears, it would not be very unreasonable to assume equality of the variances,  $\sigma_i^2$ , from composite commodity to composite commodity. In that case, the above condition would reduce to

$$\sum_{i=1}^g N_i b_i = 0. \quad \dots (6.3)$$

For small values of  $\rho_i$  the error involved may not be much. If, therefore, the composite commodities could be so taken as to ensure at least a small value for  $\rho_i$ , the practice under consideration would be commendable. This practice has a practical advantage in that it involves a lesser effort than what is required to find a price averaged with quantities sold in the market.

The illustrations cited below will show how the condition,  $\rho_i = 0$ , may be utilised with advantage in the construction of the index on minimum computation.

# PRECISION IN THE CONSTRUCTION OF COST OF LIVING INDEX NUMBERS

## 7. NUMERICAL ILLUSTRATIONS

Table 1 shows the calculation of the index on Food. The consumption of food has been divided here into 25 composite commodities. Supposing that the index on Food has been correctly calculated on these 25 composite commodities, this numerical example may be utilised to show how the composite commodities could be further grouped so that the same index on Food could be arrived at with a lesser number of composite commodities.<sup>1</sup> Although some of these suggested groupings, which have been made here on the basis of zero-correlation, may not be practicable, these groupings have, none-the-less, been shown as illustrations to point out how the above result could be exploited in computing the index on minimum effort.

TABLE 1. PRICE RELATIVES OF FOOD ARTICLES AND THEIR WEIGHTS

items	price relatives	weights (in per- centages)	items	price relatives	weights (in per- centages)
(1)	(2)	(3)	(1)	(2)	(3)
1. rice	102	27.64	14. other milk products	96	0.63
2. rice products	86	1.99	15. potato	62	4.13
3. wheat and wheat products	125	8.28	16. onions	155	0.69
4. other cereals & cereal products	95	0.72	17. other non-leafy vegetables	57	8.31
5. pulses	91	5.09	18. leafy vegetables	83	3.47
6. edible oils	72	7.93	19. fish	76	7.54
7. vegetable ghee	102	0.93	20. meat	97	1.74
8. salt	92	0.41	21. eggs	80	0.39
9. spices	85	3.93	22. fruits	107	1.17
10. sugar	93	4.67	23. tea and coffee	101	1.34
11. gur	123	0.56	24. other refreshments & sweets	90	3.20
12. milk	101	3.44	25. other food articles	116	1.09
13. butter and ghee	95	0.71			100.00

Index = 91.43

The correlation coefficient between the 25 price relatives and weights is  $-0.1390$ . The index on Food calculated from the 25 composite commodities is 91.43, while a simple arithmetic average of the price relatives is 95.28. As there is a negative correlation, the weighted index is lesser in magnitude than the simple arithmetic average of the price relatives.

The 25 price relatives for the 25 composite commodities of Food have been plotted against the respective weights in the diagram. From the graph it can be readily determined as to what of these 25 commodities could be further grouped. 5 sets of groupings have been suggested and the index for each calculated, as shown at the bottom of each set in Table 2.

<sup>1</sup> The pooling cannot be extended too far. Against this, the magnitude of the error-variance of the index estimated through the pooling may be a limiting factor.



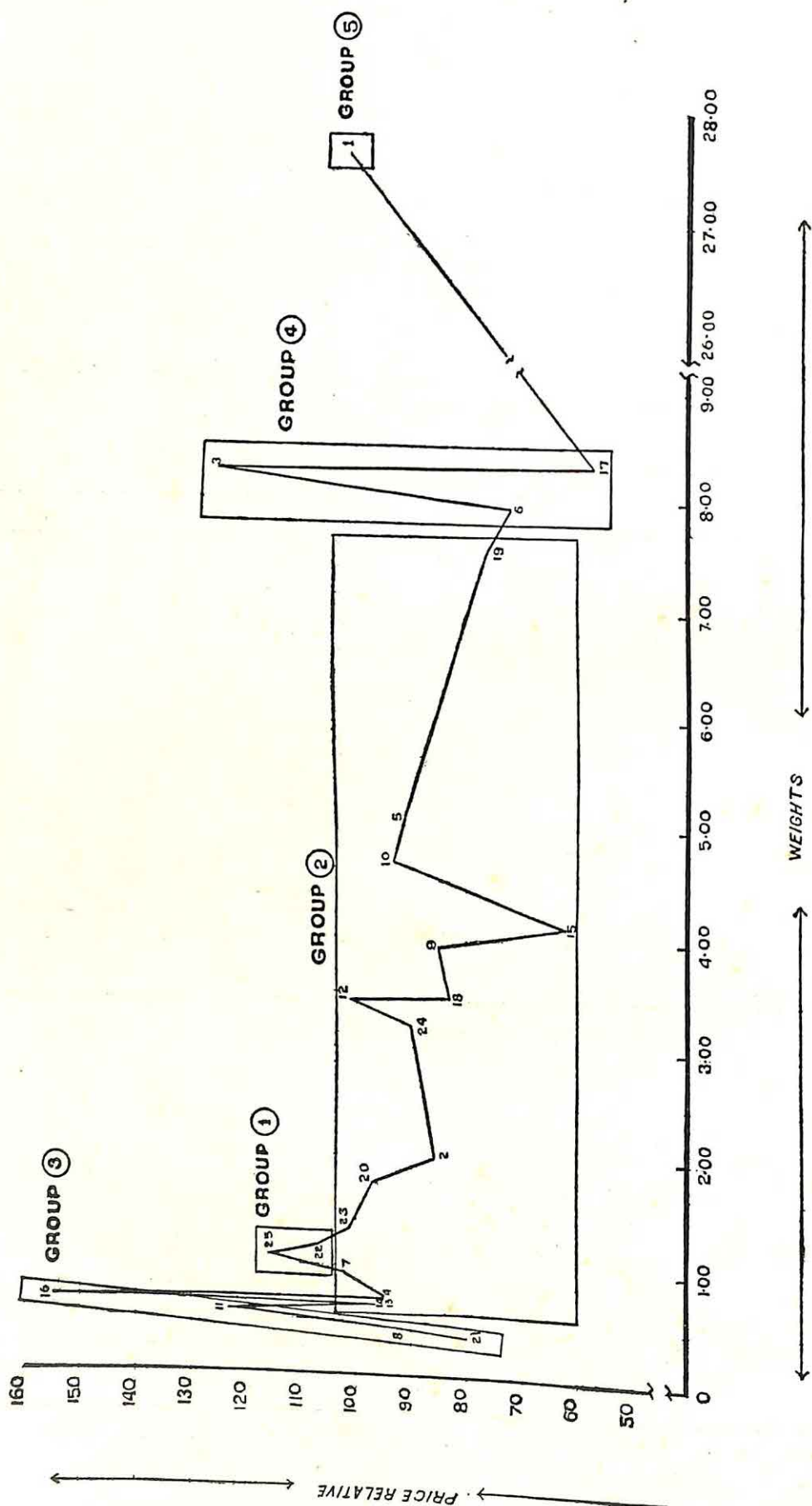


Figure 1. The grouping of consumption items

# PRECISION IN THE CONSTRUCTION OF COST OF LIVING INDEX NUMBERS

TABLE 2. DIFFERENT SETS OF GROUPS

number of groups	sets					
	I	II	III	IV	V	VI
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	25, 22	25, 22	25, 22	25, 22	25, 22, 7	25, 24, 23, 22
2	24, 23, 20, 19, 18, 15, 14, 13, 12, 10, 9, 7, 5, 4, 2	24, 20, 18, 15, 14, 13, 12, 10, 9, 5, 4, 2	24, 18, 15, 12, 10, 9, 5, 2	24, 18, 15, 12, 10, 9, 5, 2	24	21, 20, 19
3	21, 16, 11, 8	23, 7	23, 7	23, 7	23	18, 17, 16, 15
4	17, 6, 3	21, 16, 11, 8	21, 16, 11, 8	21, 16, 11, 8	21, 16, 11, 8	14, 13, 12
5	1	19	20	20	20	11, 10
6	—	17, 3	19, 6	19	19	9, 8, 7, 6
7	—	6	17, 3	17, 3	18, 15, 12, 10, 9	5
8	—	1	14, 13, 4	14, 13, 4	17, 6, 3	4, 3, 2
9	—	—	1	6	14, 13, 4	1
10	—	—	—	1	5	—
11	—	—	—	—	2	—
12	—	—	—	—	1	—
index for food	91.52	91.94	91.26	91.25	91.40	95.93

In the determination of the groups visually from the graph, use was made of the following criteria:

- (i) Equality of the weights, or
- (ii) Equality of the price relatives or
- (iii) Equality of both weights and price relatives, or
- (iv)  $\rho_i = 0$ .

The existence of criterion (iv) will be evident from the regression line being parallel to the axis of weights. Criteria (i), (ii) and (iii) are only special cases for  $\rho_i = 0$ .

The five-group-set (vide col. (2) of Table 2) has been marked out in the diagram indicating what commodities have been pooled together. The index is 91.52, while the index from the 25 commodities (groups) is 91.43. The agreement here is quite close. The agreement is also sufficiently close in the other sets except in the set shown in col. (7) of Table 2, which has been purposely suggested because such kind of grouping is usually adopted in practice. That is, items of similar nature are grouped together. The underlying presumption for this kind of grouping is perhaps the equality of the price relatives giving zero-value for the correlation coefficient. But such kind of grouping, as will be evident, may lead to



erroneous results. The calculated index is 95.93, and this is wide away from 91.43, and is nearly equal to the index which could be obtained as a simple arithmetic average of the price relatives of the 25 commodities. Adoption of a course, such as this, is, therefore, tantamount to ignoring the weights, even though the weights were relevant.

For the sake of a further illustration of the principle, a reference may be made to Table 3, where the indexes for the five conventionally accepted major groups of consumption have been shown along with the overall CLI. Let the weights of the major groups (2), (3) and (4) be pooled together, and a simple arithmetic mean taken of the three indexes (price relatives) to correspond to the pooled weight of these three major groups. These three major groups taken together will, therefore, form one major group now. In all, then, there will be three major groups instead of five. The weighted average of these three indexes (price relatives) is 95.6 which differs from the overall CLI by only 0.1.

TABLE 3. CLI FOR A MONTH IN RESPECT OF  
A TOWN FOR A SPECIFIC EXPENDITURE  
LEVEL

major groups of consumption	for the specific expenditure level		
	weight	index	weight index
(1)	(2)	(3)	(4)
1. food	58.55	91.4	5351.5
2. clothing	5.37	106.5	571.9
3. fuel & light	6.15	102.2	128.5
4. housing	9.61	100.0	961.0
5. miscellaneous	20.32	100.3	2038.1
all combined	100.00		95.5

## 8. REMARKS

The illustration cited in col.(7) of Table 2 demonstrates how non-judicious grouping, a kind of which is usually adopted in practice, might lead to serious errors. The other illustrations would point to the probable good use which could be made of the basic principles in bringing precision in the construction of index numbers.

## REFERENCES

- BANERJEE, K. S. (1956a): A note on the optimal allocation of consumption items in the construction of a cost of living index. *Econometrica*, 24, 3, 294-295.
- (1956b): A comment on the construction of price index numbers. *Applied Statistics*, 5, 3, 207-210.
- (1956c): Simplification of the derivation of Wald's formula for the cost of living index. *Econometrica*, 24, 3, 296-298.
- FRISCH, R. (1936): Annual survey of general economic theory: The problem of index numbers. *Econometrica*, 4, 1-38.
- KONÜS, A. A. (1939): The problem of the true index of the cost of living. *Econometrica*, 7, 10-29.
- WALD, A. (1939): A new formula for the index of the cost of living. *Econometrica*, 7, 319-331.

Paper received : March, 1959.



## PRICE INDEXES AND SAMPLING

By ERLAND v. HOFSTEN

*Statistical Section, National Social Welfare Board, Stockholm*

*and*

*Indian Statistical Institute, Calcutta*

**SUMMARY:** Some practical difficulties of obtaining sampling errors for price index numbers have been discussed in this note.

In recent years it has become more and more widely accepted that statistical estimates should be accompanied by error estimates, and that information about the statistical error can only be obtained, if the estimate is based on a probability sample. There is however, one exception from this general rule, namely the field of index numbers, where very little is known and stated about the precision of the computed figures. However, two authors have recently given solutions of standard errors for price index numbers. In the view of the present author a nearer scrutiny of this aspect of the problem demonstrates that there is a certain inescapable controversy and inconsistency as regards price index numbers.

The two authors are Banerjee (1956) and Adelman (1958). Banerjee points out that prices are normally collected only for a few of the items to be covered by the index. He then gives the unbiased estimate of the index as well as the variance. Banerjee's solution assumes that only two points of time 0 and 1 are compared and that Laspeyres' formula is used. This implies that prices are assumed available at period 1 for all articles which were available at period 0 and that new items are ignored.

Adelman completely overlooks Banerjee's paper, although it was published in a widely circulated journal two years earlier. She does not specify the index formula used; her index concept is one of price relatives between two periods not too far apart, and she states that the weights applied must not be out of date. For comparisons over longer intervals she arrives at the chain index solution.

To start with let us consider the problem of comparing two periods only. During the first period we have, with usual notations, certain quantities,  $q_0$ , and prices,  $p_0$ , of all articles on the market.

In order to take a probability sample we must define the universe properly and construct a frame, from which to draw the sample. If we consider *one* period only, our universe may consist of all the purchases which have taken place during the period (possibly purchases by some properly delimited population category, such as working class families, etc.). A sampling frame should consist of all these purchases; in this connection we ignore the practical difficulties to obtain such a frame.

It is important to note that not only the quantities but also the prices refer to a period and that the prices are "paid prices" not "demanded prices." This distinction is of importance, not so much because of bargaining, discounts, sales, etc., but because of the definition of the universe.



It is not quite clear which definition Irma Adelman employs. Her statement "since loss leaders and similar sub-normal price situations often exist on Thursday, Friday and Saturday, all the pricing was done during the early part of the week (p.245)" seems to imply that she uses the demanded price as definition.

For the index computation we cannot be satisfied with having a sample referring to one period only; the index implies a *comparison* between at least two periods. If we accept the Laspeyres' solution, this implies that we base the sample of items on the conditions prevailing during period 0 (= quantities purchased and prices actually paid). For period 1 we will want to ascertain the amount of money required in order to buy the same quantities in the new price situation. This implies a hypothetical question. Thus if an item is available but not at all purchased in situation 1, it will nevertheless enter into the index computation. Prices for situation 1 will not be paid prices but demanded prices, and it will not be possible for situation 1 to construct any sampling frame, which corresponds to the one in situation 0. This is not very satisfactory, but is a necessary consequence of the Laspeyres' approach.

The Paasche index, implying the reverse of the Laspeyres' index, of course does not solve the problem.

The indifference defined index also tries to give an answer to a hypothetical question, i.e., what is the amount of money required to attain an unchanged indifference level (= generally less amount of money than required for the Laspeyres' solution)? Information about the actual position of the consumer in situation 1, does not solve the problem if the index is based on the indifference level in situation 0.

It remains to be seen whether a universe can be defined where "paid prices" can be used for both periods 0 and 1. But then only articles actually bought during both periods 0 and 1 will be included, because no price relatives can be formed for items purchased only during one of the periods. And what about the weights, shall they be an average for the two periods and then why? This solution will be rather vague and unsatisfactory.

And finally, is it possible to envisage "demanded prices" as the price definition right through? Clearly not, in any case it seems difficult to find any universe then, and where do the quantities come in?

The differences between demanded prices and actually paid prices is also very clear, if we consider the problem raised in Stone (1956), when the price per unit is a function of the quantity purchased. This is often the case for electricity and telephone charges, where a basic payment is made, but also occurs regarding other items of expenditure, where bulk purchases may lead to a lower price per unit. If the Laspeyres' solution is employed, the index will not show any change, as not more money is required in order to keep the consumption pattern unaltered. But if paid prices are used, account must be taken of the price change which has been a consequence of the altered consumption.

The problems discussed above refer to the fact that the universe is changing. Such changes are most marked as regards clothing and so-called miscellaneous items, whereas they are less marked for food items. Incidentally most authors on index numbers overlook these problems, because they choose examples among food items. However, looking upon the budget as a whole, the problems of the changing universe are severe; for evidence see Hofsten (1952).



## PRICE INDEXES AND SAMPLING

If the periods compared are near each other, it is tempting to state that then the changes of the universe must be so small that they can be overlooked. In order to make possible comparisons over long intervals, we must then resort to the chain index solution.

The chain index implies an integral solution of the index problem [cf. Divisia (1925).] If this solution is chosen, then *it is necessary that the infinitesimal expression for the index, i.e. the index for each separate link, is correct.* If this is not the case, the chain index only implies a comfortable technique, by which the intrinsic problems of comparing two distant periods are avoided. Incidentally the chain index solution is not available for geographical comparisons, at least not between different countries.

Adelman states about the chain index that "the ease with which new products or qualities can be incorporated into our scheme (and obsolete items eliminated) provides a significant advantage over the current system" (p.243). This advantage, in my mind implies a great danger, because it violates the principle that the infinitesimal expression for the index must be correct.

There is one additional problem of a partly practical character. A computation of a standard error for an index will in the first hand refer to a comparison between two periods only [as in Banerjee (1956) and Adelman (1958)]. But in actual practice indexes are most often given in the form of long regular series. If the series is computed as a chain index, what standard error formula shall then be used? And as the consumer will desire to compare any single index figure with any other figure, what standard errors shall be given?

My conclusion from the above arguments is that there is no such thing as a statistical precision for a price index. Attempts to define the index in a statistical way, applying modern theory of sampling, only demonstrate that there is no satisfactory solution available. We may, therefore, just as well keep to the old practice and define the price index in an operational way and abstain from giving standard errors. This, of course, does not exclude the usefulness of applying the chain index solution or of basing the selection of items on probability sampling and making analyses of the precision of price measurements. But when applying the chain index solution we must not allow the substitution of some items against others without making quality adjustments; see Hofsten (1952) and Stone (1956).

### REFERENCES

- ADELMAN, IRMA (1958): A new approach to the construction of index numbers. *The Review of Economics and Statistics*, **XL**, No. 3, 240-249.
- BANERJEE, K. S. (1956): A Note on the optimal allocation of consumptive items in the construction of cost of living index. *Econometrica*, **24**, 294-295.
- DIVISIA, FRANCOIS (1925): L'indice monetaire et la theorie de la monnaie. *Revue d'economie politique*, Paris.
- HOFSTEN, E. V., (1952): *Price Indexes and Quality Changes*, Allen & Union, Stockholm and London.
- STONE, R., (1956): *Quantity and Price Indexes in National Accounts*. OEEC, Paris.

*Paper received : April, 1959.*



## CORRIGENDA

**Bias in Estimation of Serial Correlation Coefficients :** By A. Sree Rama Sastry, *Sankhyā*, **11**, 281-296.

Formula (11) on page 283 should be read

$$g_k = - \frac{\sum_{i=1}^{T-k-1} (T-k-i) \left\{ \mu_2(k+i) + \mu_2(|k-i|) - \frac{2\mu_2(k)\mu_2(i)}{\mu_2(0)} \right\}}{(T-k-1)(T-k)\mu_2(0) - 2 \sum_{i=1}^{T-k-1} (T-k-i)\mu_2(i)} \quad \dots (11)$$

The author wishes to thank Mr. E. G. Phadia for pointing out the printing error.

**Expressions for The Lower Bound to Confidence Coefficients :** By Saibal Kumar Banerjee, *Sankhyā*, **21**, 127-140.

1.  $\frac{t\hat{\lambda}}{n}$  occurring in (i) expression (2.4.3), (ii) first sentence of para 2.5 and (iii) table heading of Table 1, all at page 129, should be read as  $t\sqrt{\frac{\hat{\lambda}}{n}}$
2.  $B_2$  occurring in para 3.7, page 132, is

$$\bar{B}_2 = \frac{\sum_{i=1}^k \lambda_i^2 B_{2i}}{\sum_{i=1}^k \lambda_i^2}$$

## ACKNOWLEDGEMENT

Thanks are due to the following for their kind help in the editing of the papers published in this issue.

Dr. R. R. Bahadur

Dr. K. S. Banerjee

Dr. D. Basu

Dr. I. M. Chakravarti

Dr. E. v. Hofsten

Professor D. B. Lahiri

Dr. A. Matthai

Dr. Sujit Kumar Mitra

Dr. C. R. Rao

Dr. J. Roy



















